McGRAW-HILL
ENCYCLOPEDIA OF
SCIENCE & TECHNOLOGY

9 I - LEV

www.MHEST.com

I

I–spin — Ixodides

I-spin

A quantum-mechanical variable or quantum number applied to quarks and their compounds, the strongly interacting fundamental hadrons, and the compounds of those hadrons (such as nuclear states) to facilitate consideration of the consequences of the charge independence of the strong (nuclear) forces. This variable is also known as isotopic spin, isobaric spin, and isospin.

The many strongly interacting particles (hadrons) and the compounds of these particles, such as nuclei, are observed to form sets or multiplets such that the members of the multiplet differ in their electric charge and magnetic moments, and other electromagnetic properties but are otherwise almost identical. For example, the neutron and proton, with electric charges that are zero and plus one (in units of the magnitude of the electron charge), form a set of two such states. The pions, one with a unit of positive charge, one with zero charge, and one with a unit of negative charge, form a set of three. It appears that if the effects of electromagnetic forces and the closely related weak nuclear forces (responsible for beta decay) are neglected, leaving only the strong forces effective, the different members of such a multiplet are equivalent and cannot be distinguished in their strong interactions. The strong interactions are thus independent of the different electric charges held by different members of the set; they are charge-independent. *See* ELEMENTARY PARTICLE; FUNDAMENTAL INTERACTIONS; HADRON; STRONG NUCLEAR INTERACTIONS.

The I-spin (I) of such a set or multiplet of equivalent states is defined such that Eq. (1) is satisfied,

$$N = 2I + 1 \qquad (1)$$

where N is the number of states in the set. Another quantum number I_3, called the third component of I-spin, is used to differentiate the numbers of a multiplet where the values of I_3 vary from $+I$ to $-I$ in units of one. The charge Q of a state and the value of I_3 for this state are connected by the Gell-Mann-Okubo relation, Eq. (2), where Y, the charge offset, is called

$$Q = I_3 + \frac{Y}{2} \qquad (2)$$

hypercharge. For nuclear states, Y is simply the number of nucleons. Electric charge is conserved in all interactions; Y is observed to be conserved by the strong forces so that I_3 is conserved in the course of interactions mediated by the strong forces. *See* HYPERCHARGE.

Similarity to spin. This description of a multiplet of states with I-spin is similar to the quantum-mechanical description of a particle with a total angular momentum or spin of j (in units of $\hbar$, Planck's constant divided by 2π). Such a particle can be considered as a set of states which differ in their orientation or component of spin j_z in a z direction of quantization. There are $2j + 1$ such states, where j_z varies from $-j$ to $+j$ in steps of one unit. To the extent that the local universe is isotropic (or there are no external forces on the states that depend upon direction), the components of angular momentum in any direction are conserved, and states with different values of j_z are dynamically equivalent.

There is then a logical or mathematical equivalence between the descriptions of (1) a multiplet of states of definite I and different values of I_3 with respect to charge-independent forces and (2) a multiplet of states of a particle with a definite spin j and different values of j_z with respect to direction-independent forces. In each case, the members of the multiplet with different values of the conserved quantity I_3 on the one hand and j_z on the other are dynamically equivalent; that is, they are indistinguishable by any application of the forces in question. *See* ANGULAR MOMENTUM; SPIN (QUANTUM MECHANICS).

Relative intensities of virtual transitions determined by isobaric spin symmetry

Transition	Relative intensity
$p \rightarrow n + \pi^+$	2/3
$p \rightarrow p + \pi^0$	1/3
$n \rightarrow n + \pi^0$	1/3
$n \rightarrow p + \pi^-$	2/3

Importance in reactions and decays. The charge independence of the strong interactions has important consequences, defining the intensity ratios of different charge states produced in those particle reactions and decays which are mediated by the strong interactions. A simple illustration is provided by the virtual transitions of a nucleon to a nucleon plus a pion, transitions which are of dominant importance in any consideration of nuclear forces. The neutron and proton form a nucleon I-spin doublet with $I = {}^1/_2$; the pions form an I-spin triplet with $I = 1$. If initially there are one neutron and one proton, with no bias in charge state or I_3, and the strong forces responsible for the virtual transitions do not discriminate between states with different charge or different values of I_3, then it follows that in the final system there must be equal probabilities of finding each charge member of the pion triplet and equal probabilities of finding a neutron or proton. This condition, that the strong interactions cannot differentiate among the members of an isobaric spin multiplet, determines the relative intensities of the transitions (see **table**).

Using the same kind of argument, it is easy to see that the conditions of equal intensity of each member of a multiplet cannot be fulfilled in a transition from an initial doublet to a final state of a doublet and a quartet. Therefore, none of the individual transitions is allowed by charge independence, though charge or I_3 is conserved in the decays. In general, decays are allowed for a transition A → B + C only if inequality (3) is satisfied. This is analogous to the

$$|I(B) + I(C)| \geq |I(A)| \geq |I(B) - I(C)| \qquad (3)$$

vector addition rule for spin or angular momentum; the strong interactions conserve I-spin in the same manner as angular momentum is conserved. *See* SELECTION RULES (PHYSICS).

Classification of states. I-spin considerations provide insight into the total energies or masses of nuclear and particle states. The fundamental constituents of nuclei are the nucleons, the neutron and proton–spin-${}^1/_2$ fermions which must obey the Pauli exclusion principle to the effect that the sign of the wave function that describes a set of identical fermions must change sign upon exchange of any two fermions. Similarly, hadrons are described as compounds of quarks, which are also spin-${}^1/_2$ fermions. The two fermions that make up an I-spin doublet can be considered as different charge states of a basic fermion, even as states with the spin in the plus and minus direction of quantization are considered as different spin states of the fermion. The extended Pauli exclusion principle then requires that the wave function amplitude change sign upon exchange of spin direction, charge, and spatial coordinates for two (otherwise) identical fermions. *See* EXCLUSION PRINCIPLE.

A space state $u(r)$ of two identical fermions, where r is the vector distance between the two particles, will be even upon exchange of the two particles if $u(r)$ has an even number of nodes, and will be odd under exchange if there is an odd number of nodes. With more nodes, the space wavelength is smaller, and the momentum and energy of the particles are larger. The lowest energy state must then have no spatial nodes and must be even under spatial interchange. From the Pauli principle, the whole wave function must be odd, and then the exchange under spin and I−spin coordinates must be odd. Using this kind of argument, the low-mass (low-energy) states of light nuclei can be classified in terms of their I-spin symmetries.

An application of the same principle, that the space wave function of the lowest state must be even under exchange, was an important element in the discovery of a new quantum number (labeled color) for quarks, the elementary constituent of the hadrons, and concomitantly the unfolding of a deeper understanding of fundamental particles.

Basis for charge independence. The basis for the symmetry described by I-spin is to be found in the quark structure of the strongly interacting particles and the character of the interactions between quarks. All of the strongly interacting particles are quark compounds; the conserved fermions, such as the neutron and proton, are made up of three quarks; the bosons, such as the π mesons, are quark-antiquark pairs. There are six significantly different quarks arranged in three pairs of charge $+{}^2/_3$ and charge $-{}^1/_3$ particles, ($u^{2/3}$, $d^{-1/3}$), ($c^{2/3}$, $s^{-1/3}$), and ($t^{2/3}$, $b^{-1/3}$), called up and down, charm and strange, and, top and bottom. The quarks interact through their strong interacting color charges that couple the quarks to gluons in a manner analogous to the coupling of electrical charge to photons. Rather than one kind of electrical charge, and its negative, quarks have three kinds of color charges (conventionally labeled r for red, y for yellow, and b for blue), and their negatives. Even as there are three different colors, there are $3 \times 3 - 1 = 8$ different gluons labeled with color and anticolor (with the so-called white combination ruled out), rather than the one photon that carries the electromagnetic forces.

Since each kind of quark carries exactly the same sets of color charge, the strong forces between two quarks are exactly the same as the forces between two other quarks. However, the simple consequences of this color independence of the strong forces are largely obviated by the fact that the six quarks have different masses. Hence, the Σ^+ hyperon, (uus), is appreciably heavier than the proton, (uud), even as the s-quark is heavier than the d-quark, though the quark-quark forces holding the systems together are the same.

However, the masses of the two lightest quarks, the *u* and *d*, are almost the same, differing by only a few electron masses. Then, for the many situations that this small difference can be neglected, the *u* and *d* quarks cannot be differentiated by the strong forces; that is, the neutron (*duu*) and proton (*ddu*) are identical under the strong forces, which is just the symmetry described by I-spin.

There are some effects where the mass difference, equal to about two pion masses, between the *s* and *u* quark, or the *s* and *d* quark, can be neglected, or for which corrections can be calculated. For such effects twofold symmetries like I—spin, called, respectively, V-spin and U-spin, are useful. In principle, similar broken symmetries exist for compounds of the heavier quarks, but the symmetry breaking that follows from their much larger mass differences so obscures the underlying symmetry that it is not useful. *See* COLOR (QUANTUM MECHANICS); GLUONS; QUANTUM CHROMODYNAMICS; QUARKS; SYMMETRY LAWS (PHYSICS). Robert K. Adair

Bibliography. F. Halzen and A. Martin, *Quarks and Leptons*, 1984; N. A. Jelley, *Fundamentals of Nuclear Physics*, 1990; G. Kane, *Modern Elementary Particle Physics*, updated ed., 1993; L. B. Okun, *Leptons and Quarks*, 1980, paper 1985; S. S. Wong, *Introductory Nuclear Physics*, 2d ed., 1999.

Ice cream

A commercial dairy food made by freezing while stirring a pasteurized mix of suitable ingredients. The product may include milk fat, nonfat milk solids, or milk-derived ingredients; other ingredients may include corn syrup, water, flavoring, egg products, stabilizers, emulsifiers, and other non-milk-derived ingredients. Air incorporated during the freezing process is also an important component.

The structure of ice cream is complex. It consists of solid, gaseous, and liquid phases; ice crystals and air cells are dispersed throughout the liquid phase, which also contains fat globules, milk proteins, and other materials.

Composition. In the United States, ice cream composition is regulated by Federal Frozen Dessert Standards of Identity, which set forth minimum composition requirements of not less than 10% milk fat and not less than 20% total milk solids. Ice cream must weigh not less than 4.5 lb/gal (0.54 kg/liter) and must contain not less than 1.6 lb food solids/gal (0.26 kg/liter). In the case of bulky flavors, the fat may be not less than 8%, nor can the total milk solids be less than 16% in the finished food. Ingredient and nutritional-requirements labeling is included in the standards.

The composition of ice cream may vary depending on whether it is an economy brand satisfying minimal requirements, a trade brand of average composition, or a premium brand of superior composition. The components by weight of an average-composition ice cream are 12% fat, 11% nonfat milk solids, 15% sugar, and 0.3% vegetable gum stabilizer.

An average serving of 4 fl oz (120 ml) of vanilla ice cream with 10% milk fat provides about 135 calories, 0.4 oz (11.2 g) fat, 0.8 oz (22.4 g) protein, 0.56 oz (15.9 g) carbohydrate, 0.003 oz (88 mg) calcium, 0.0024 oz (67 mg) phosphorus, 270 International Units vitamin A, and 0.0000058 oz (0.164 mg) riboflavin. Thus, a serving of ice cream provides more than 10% of the U.S. Recommended Daily Dietary Allowance (USRDA) for riboflavin and calcium and about 5% of the USRDA for protein for adults and teenagers. *See* NUTRITION.

French ice cream may contain a relatively high fat content, have a slight yellow color, and contain egg yolk solids. Both this product and ice cream carry the same requirements. French ice cream must contain not less than 1.4% egg yolk solids for plain flavor and not less than 1.12% for bulky flavors. It is usually sold as a soft-serve product but may be sold as a prepackaged hardened product. Gelato is Italian for ice cream; it commonly increases in volume by only one-third on freezing.

The composition of ice milk is similar to that of ice cream except that it must contain more than 2% but not more than 7% milk fat, not less than 11% total milk solids, and not less than 1.3 lb food solids/gal (0.15 kg/liter).

Sherbet is made with about 1 part ice cream mix to 4 parts water ice mix and is manufactured in the same way as ice cream. It must weigh not less than 6 lb/gal (0.7 kg/liter). Sherbet contains not less than 1% or more than 2% fat, and total milk solids of not less than 2% or more than 5% by weight of the finished food; it must also contain 0.35% edible citric or natural fruit acid. A sherbet made with addition of egg yolk is known as frappe.

Frozen dairy desserts. Products classed as frozen dairy desserts include French ice cream, frozen custard, ice milk, sherbet, water ice, frozen dairy confection, dietary frozen dairy dessert, and mellorine (imitation ice cream). Most of these desserts can be sold in the soft-serve form, as they are when first removed from the freezer at about 20°F (−7°C), or in the hard form as they are when the temperatures are reduced to about 8°F (−13°C). A frozen dessert mix is the combination of ingredients, usually in the liquid form and ready to be frozen. Dry mixes must be dissolved in water before freezing. Unflavored mixes may be flavored before freezing, as with vanilla and chocolate, or may have flavorings added to the soft frozen product as it exits the freezer, as with fruit, nut, and candy type desserts.

Water ice is made from water, sugar, corn sweeteners, fruit juices, flavorings, and stabilizers. It contains no milk solids. Sorbet is similar; in addition it contains finely divided fruit and is whipped while being frozen. There is no federal standard for sorbet in the United States.

Dietary frozen desserts are comparatively low in calories. Ice milk commonly contains about one-third the fat of ice cream. Proposed U.S. Federal Standards permit 2–5% fat in ice milk and establish a category of light (lite) ice cream with 5–7% fat. To provide frozen desserts for diabetics, sugar (sucrose) is replaced with nonnutritive sweeteners, and complex carbohydrates, such as sorbitol, are added to provide desirable texture.

Mellorine is manufactured in a process similar to that for ice cream, but is not limited to milk-derived nonfat solids and may be composed of animal or vegetable fat or both, only part of which may be milk fat. It contains not less than 6% fat and not less than 2.7% protein by weight.

Frozen dairy confections are produced in the form of individual servings and are commonly referred to as novelties, including bars, stick items, and sandwiches. These constitute an important segment of the industry.

Soft frozen dairy products make up about 9.6% of the total frozen desserts produced in the United States. Per capita consumption is 5 qt (4.7 liters) per year. The soft frozen product is usually drawn from the freezer at 18 to 20°F (−8 to −7°C) and is served directly to the consumer. Ice milk accounts for about three-fourths of the soft-serve products that are manufactured in the United States. These products usually vary from 2 to 7% fat, 11 to 14% nonfat milk solids, 13 to 18% sugar, and have about 0.4% stabilizer. Frozen custard and ice cream are also important soft-serve products.

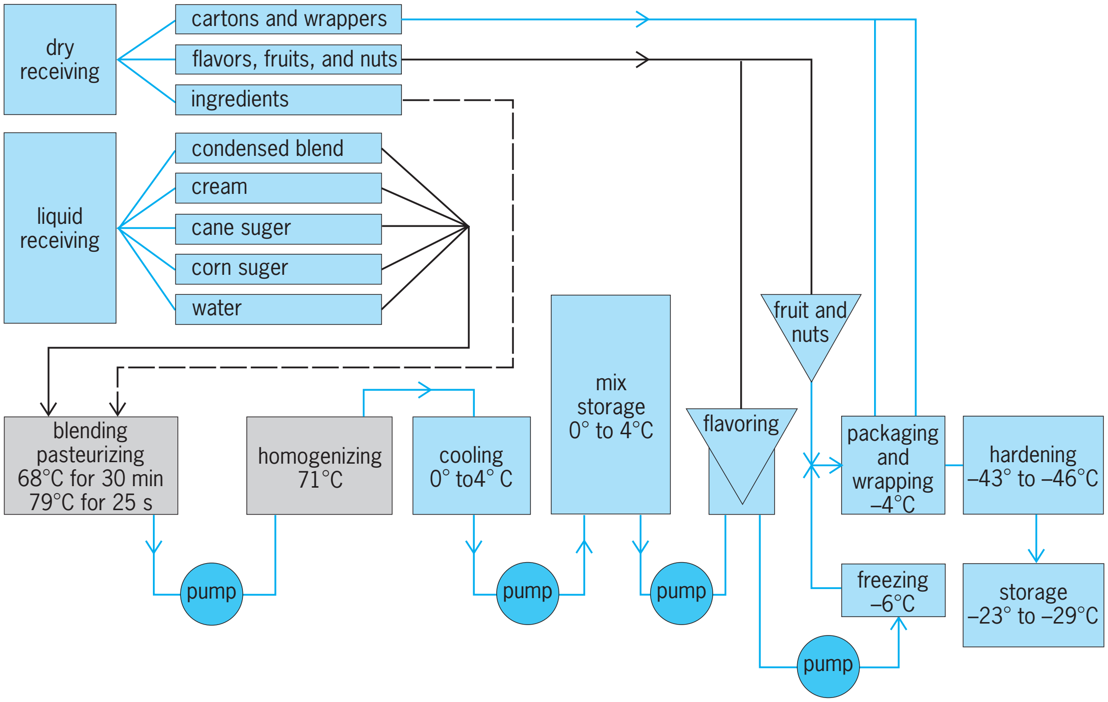
Commercial manufacture. In ice cream manufacture the basic ingredients are blended together (see **illus.**). The process ranges from small-batch operations, in which the ingredients are weighed or measured by hand, to large automated operations, where the ingredients, all in liquid form, are metered into the mix-making equipment. The liquid materials, including milk, cream, concentrated milk, liquid sugar syrup, and water, are mixed. The dry solids, such as nonfat dry milk, dried egg yolk, stabilizer, and emulsifier, are blended with the liquid ingredients. This liquid blend is known as the mix. Following the blending operation, the mix is pasteurized, homogenized, cooled, and aged. Pasteurization destroys all harmful microorganisms and improves the storage properties of the ice cream. *See* PASTEURIZATION.

The hot mix is pumped from the pasteurizer through a homogenizer, which reduces the fat globules to less than 2 micrometers in diameter. Homogenization is accomplished by forcing the mix through the homogenization valves under pressure in the range 1500–3000 lb/in.^{2} (103–207 kilopascals). This phase of the manufacturing process results in a uniform, smooth product and limits churning of the fat during the freezing process.

The mix is cooled immediately to 32–40°F (0–4°C); then it is aged 4–12 h in order to permit development of favorable hydration and particle orientation effects, which serve to improve the body and texture in the finished ice cream.

Soluble flavoring materials are usually added to the mix just before the freezing process, but fruits, nuts, and candies are not added until the ice cream is discharged from the freezer. In the United States, popular flavors of ice cream include vanilla, chocolate, strawberry, butter pecan, and "cookies 'n cream" (broken cookies are blended into the ice cream), and there are many other flavors. Neapolitan is one-third each of vanilla, chocolate, and strawberry extruded simultaneously into one container.

Ice cream is frozen in batch or continuous



Flow chart for ice cream processing. °F = (°C × 1.8) + 32. (*Arbuckle and Co.*)

freezers. The continuous freezing process is more effective, as it freezes more rapidly and produces a product with finer texture. Continuous freezers vary in capacity from 85 gal (323 liters) to more than 2000 gal (7600 liters) per hour.

During freezing, air is incorporated into the mix, resulting in increased volume. This increase in volume is called overrun. The drawing temperature of the ice cream from the freezer is about 21°F (−6°C). The ice cream is packaged at this temperature and cooled further to about −20°F (−29°C) in a hardening room, where it is stored until marketed.

The ice cream industry of the United States has been described as a high-volume, highly automated, progressive, very competitive industry composed of large and small factories. It has developed on the basis of abundant and economical supply of ingredients. Also contributing to the growth of the industry are factors such as achievement of high quality; favorable consumer response; excellent transportation, refrigeration, and distribution; advanced packaging; and abundant home refrigeration.

The annual per capita production in the United States is about 15 qt (14 liters) of ice cream. Combined production of ice cream, ice milk, sherbet, and mellorine amounts to about 22 qt (21 liters) per person. This amounts to 1.32 billion gal (5 billion liters) per year of ice cream and related products.

Wendell S. Arbuckle; Robert T. Marshall

Sources of microorganisms. The cream, condensed- and dry-milk products, egg products, sugar, stabilizers, flavoring and coloring materials, and fruits and nuts used as ingredients in ice cream may be contaminated with microorganisms. Therefore the ice cream mix must be pasteurized before freezing because a low temperature in itself is not sufficient to kill microorganisms. Other possible sources of microorganisms are equipment and utensils used in manufacture, scoops and dippers, vendors, employees, and air. Stringent regulations must be established for all manipulations of ice cream from producer to consumer in order to avoid recontamination, especially with regard to pathogens. Once incorporated, microorganisms survive a long time in ice cream.

Kinds of microorganisms. A wide variety of microorganisms may be found, including streptococci, micrococci, coliform bacteria, spore-formers, yeasts, and molds. No selection of species by freezing takes place. Since no growth and only slow destruction of microorganisms in the final product occur, the flora of ice cream consists of organisms surviving pasteurization or incorporated by contamination. With well-controlled sanitation during production, plate counts may be as low as a few hundred; with careless methods they may be in the millions. Normally a slight increase of the bacterial count takes place during manufacture, especially after homogenization and freezing, due to the breaking up of clumps of bacteria. A plate count not greater than 50,000–100,000 per gram is regarded as acceptable from a sanitary point of view in most states.

Hygienic measures. These include proper pasteurization of the mix and adequate sanitizing of equipment. Pasteurization must be more intensive with the mix than with milk because increased milk solids and sugar protect microorganisms to some degree. At least 155°F (68.4°C) for 30 min or 175°F (79.5°C) for 25 s is applied to obtain a sufficient margin of safety. In this way all pathogens are killed and total bacterial counts of 10,000 or less per gram are usually reached. Higher counts after pasteurization are due to thermoduric organisms. The source of these organisms is inadequately cleaned farm or plant equipment. Sanitizing of equipment is achieved by cleaning and disinfecting operations with heat or chemicals. *See* DAIRY MACHINERY; STERILIZATION.

Bacteriological control. For determining the hygienic quality of ice cream, the laboratory tests used are similar to those employed for milk. Of these, the standard plate count and the test for coliform bacteria have found widespread application. *See* FOOD MANUFACTURING; MILK.

William C. Winder

Bibliography. W. S. Arbuckle, *Ice Cream*, 5th ed., 1996; International Ice Cream Association, *The Latest Scoop*: *Facts and Figures on Ice Cream and Related Products*, annually; R. K. Robinson (ed.), *Modern Dairy Technology*, 2d ed., 1993.

Ice field

A network of interconnected glaciers or ice streams, with common source area or areas, in contrast to ice sheets and ice caps. The German word *Eisstromnetz*, which translates literally to ice-stream net, is sometimes used for glacial systems of moderate size (such as less than 3000 mi^2 or 8000 km^2) and is most applicable to mountainous regions. Being generally associated with terrane of substantial relief, ice-field glaciers are mostly of the broad-basin, cirque, and mountain-valley type. Thus, different sections of an ice field are often separated by linear ranges, bedrock ridges, and nunataks.

Contrast with ice sheet. An ice sheet is a broad, cakelike glacial mass with a relatively flat surface and gentle relief. Ice sheets are not confined or controlled by valley topography and usually cover broad topographic features such as a continental plateau (for example, much of the Antarctic ice sheet), or a lowland polar archipelago (such as the Greenland ice sheet). Although ice sheets are generally of very large dimension, in some regions small, rather flat ice bodies have been called ice sheets because they are thinned remnants of once large masses of this form. Small ice sheets and even ice fields are sometimes incorrectly referred to as ice caps, even though their configurations have been well characterized.

Contrast with ice cap. Ice caps are properly defined as domelike glacial masses, usually at high elevation. They may, for example, make up the central nourishment area of an ice field at the crest of a mountain range, or they may exist in isolated positions as separate glacial units in themselves. The latter type is characterized by a distinctly convex summit dome, bordered by contiguous glacial slopes with relatively regular margins not dissected by outlet valleys or abutment ridges.

Similarities and gradations. There are all gradations between ice caps, ice fields, and ice sheets. Over a period of time, a morphogenetic gradational sequence may also develop in any one region. Major ice sheets, for example, probably originate from the thickening and expansion of ice fields and the coalescence of bordering piedmont glaciers. Conversely, ice fields can develop through the thinning and retraction of a large ice sheet overlying mountainous terrane. *See* GLACIATED TERRAIN; GLACIOLOGY.

Maynard M. Miller

Ice manufacture

Commercial production of manufactured artificial ice from water, or of dry ice from the solidification of carbon dioxide, by mechanical refrigeration. Of great economic importance was the manufacture of water ice for use in refrigeration units, fishing boats, fish- and meat-packing plants, and dairies. However, there was a sharp increase in the trend toward cubing, sizing, and packaging of ice. The domestic ice market practically vanished with development of the automatic household refrigerator. *See* DRY ICE.

Most ice is made in galvanized cans that are partially immersed in brine in an ice-making tank (**Fig. 1**). Brine made of sodium chloride or calcium chloride is used. The brine is cooled by ammonia as the refrigerant evaporates in pipe coils or brine coolers submerged in the ice tank. The ice cans are filled with raw water, or treated water if it initially contains large amounts of impurities. The water is usually precooled. Cold brine circulates around the ice cans to freeze the water. Commercial ice is frozen in 300- or 400-lb (135- or 180-kg) blocks for which the freezing time with brine at 12°F (−11°C) is 38–42 h. Freezing time depends largely on the thickness of the ice cake and the brine temperature. In large plants, cans are arranged in group frames for harvesting as many as 34 cans at a time. A traveling crane picks up a frozen group, transports and drops it into a dip tank for thawing, then moves it to a can dump where the ice cakes slide out and into a storage room. The empty cans are refilled with fresh water and are returned to the ice tank by the crane.

If clear ice is required, the water in each can must be agitated with air during freezing; otherwise opaque ice is formed. Because the water must be cooled to 32°F (0°C) before freezing can start and because of system losses, about 1.6 tons (1.5 metric tons) of refrigeration is required to make 1 ton (0.9 metric ton) of ice when the raw water enters at 70°F (21°C).

The manufacture of ice in slush, flake, or cube form by continuous ice-makers (**Fig. 2**) constitutes a large portion of the commercial market. Also, small, fully automatic, self-contained, cube and flake ice-makers are used in bars, restaurants, hotels, and hospitals.

Ice is also frozen artificially in a flat horizontal sheet for ice-skating. Such skating rinks may be indoor or outdoor, permanent or portable. The rink floor is covered with pipe coils through which cold brine is circulated and over which water is sprayed

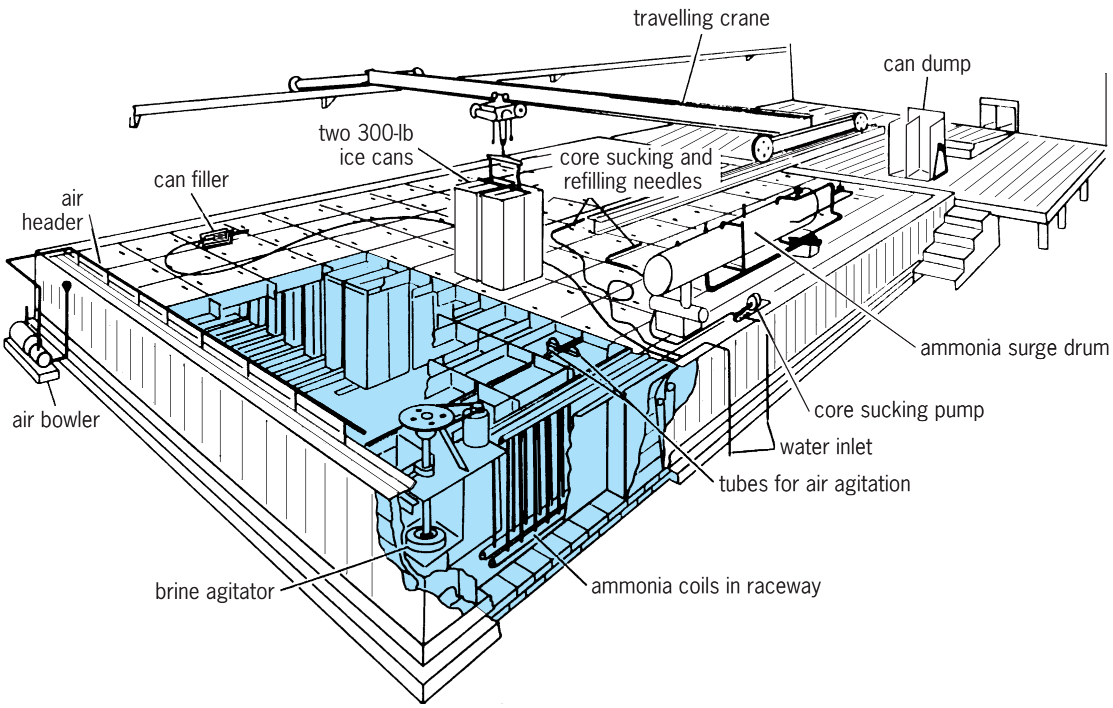


Fig. 1. Typical can ice plant. 300 lb = 135 kg. (*Worthington Pump and Machinery Corp.*)

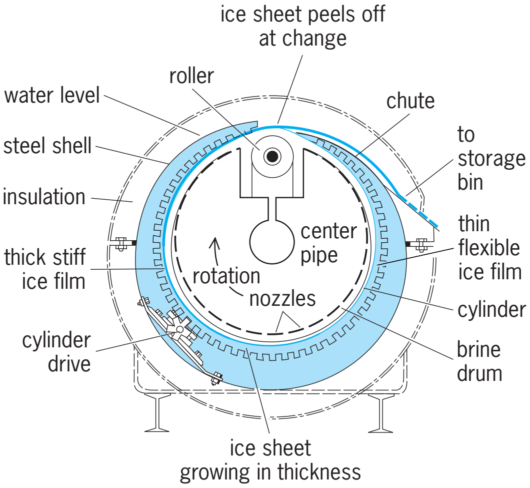


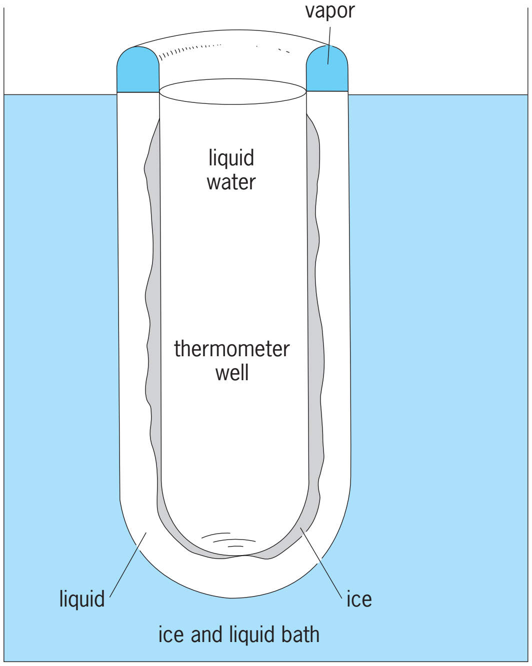
Fig. 2. Commercial flake ice machine.

until a frozen sheet $^{1}/_{2}$–$^{3}/_{4}$ in. (13–19 mm) thick is obtained. The brine is cooled in coolers by a refrigerant such as ammonia or Freon-12. *See* REFRIGERATION.

Carl F. Kayan

Ice point

The temperature at which liquid and solid water are in equilibrium under atmospheric pressure. The ice point is by far the most important "fixed point" for defining temperature scales and for calibrating thermometers. It is 273.15 K (0°C or 32°F). A closely related point is the triple point, where liquid, solid, and gaseous water are in equilibrium. It is 0.01 K (0.018°F) higher on the Kelvin scale than the ice point. The triple point has gained favor as the primary standard since it can be attained with great accuracy in a simple closed vessel, isolated from the atmosphere. Readings are reproducible to about 0.0001 K (0.0002°F), but dissolved gases or other foreign matter may raise the error to 0.001 K (0.002°F) or more. *See* TEMPERATURE; TRIPLE POINT.



Arrangement for determining triple point.

The triple-point apparatus shown in the **illustration** consists of a thermometer well that is filled with liquid water and jacketed by a cavity containing the three phases of water under a pressure of about 0.006 atm (0.09 lb/in.^{2} or 600 pascals). The ice, initially deposited by prechilling the well, melts during heat transfer from the thermometer. *See* PHASE EQUILIBRIUM.

Ralph A. Burton

Bibliography. J. F. Schooley (ed.), *Temperature: Its Measurement and Control in Science and Industry*, 1982; K. Wark, *Thermodynamics*, 6th ed., 2000.

Iceberg

A large mass of glacial ice broken off and drifted from parent glaciers or ice shelves along polar seas. Icebergs are distinguished from polar pack ice, which is sea ice, and from frozen seawater, whose rafted or hummocked fragments may resemble small icebergs. *See* GLACIOLOGY; SEA ICE.

Characteristics and types. The continental or island icecaps of both the Arctic and Antarctic regions produce icebergs where the icecaps extend to the sea as glaciers or ice shelves. The "calving" of a large iceberg is one of nature's great spectacles, considering that a Greenland iceberg may weigh over 5 million tons and that Antarctic icebergs are many times larger. An iceberg consists of glacial ice, which is compressed snow, having a variable specific gravity that averages about 0.89 gram per cubic centimeter. An iceberg's above-water volume is about one-eighth to one-seventh of the entire mass. However, spires and peaks of an eroded or weathered iceberg results in height-to-depth ratios of between 1:6 and 1:3. Tritium age experiments on melted ice from Greenland icebergs indicate that they may be around 50,000 years old. Minute air bubbles entrapped in glacial ice impart a snow-white color to it and cause it to effervesce when immersed. *See* SEAWATER; TRITIUM.

Icebergs are classified by shape and size. Terms describing icebergs include arched, blocky, domed, pinnacled, tabular, valley, and weathered, as well as bergy-bit and growler for iceberg fragments that are smaller than cottage size above water. An iceberg may last indefinitely in cold polar waters, eroding only slightly during the summer months. But an iceberg that drifts into warmer water disintegrates in a matter of weeks in sea temperatures between 40 and 50°F (4 and 10°C) or in days in sea temperatures over 50°F. A notable feature of icebergs is their long and distant drift, under the influence of ocean currents, into shipping lanes where they become navigation hazards. The normal extent of iceberg drift is shown in **Fig. 1**.

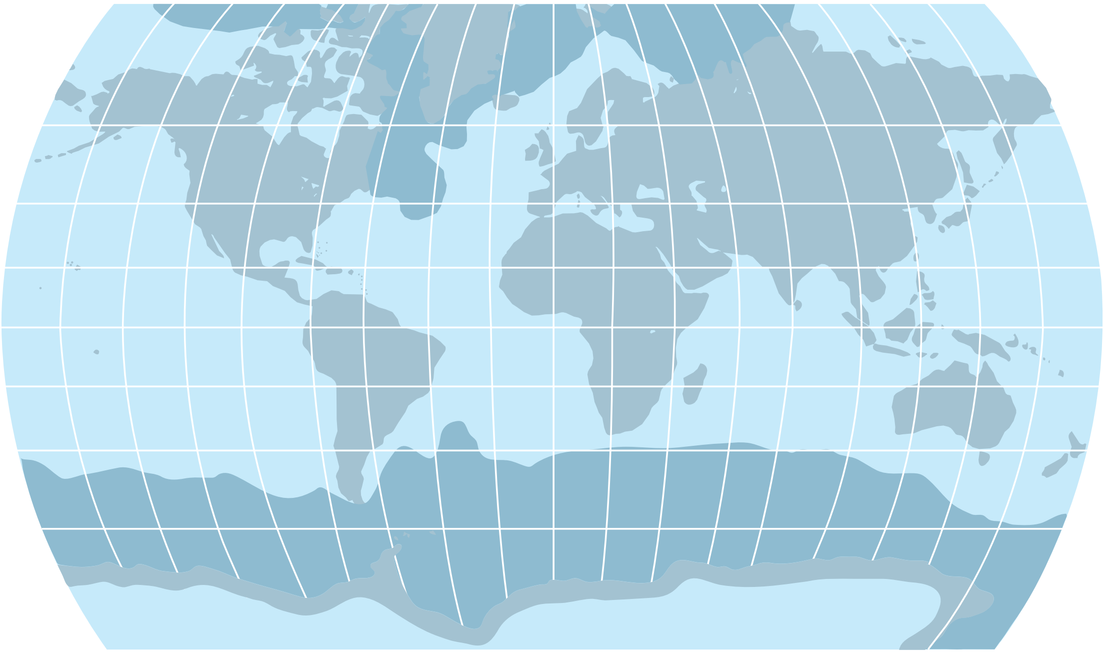


Fig. 1. Normal extent of iceberg drift, indicated by darker color tint.

Arctic icebergs. In the Arctic, icebergs originate chiefly from glaciers along Greenland's coasts. It is estimated that a total of about 16,000 icebergs are calved annually in the Northern Hemisphere, of which over 90% are of Greenland origin. But only about half of these have a size or source location to enable them to achieve any significant drift. The majority of the latter stem from some 20 glaciers along the west coast of Greenland between the 65th and 80th parallels of latitude. The most productive glacier is the Jacobshavn at latitude 68°N, calving about 1400 icebergs yearly, and the largest is the Humboldt Glacier at latitude 79° with a seaward front extending 65 mi (104 km). The remainder of the Arctic icebergs come from East Greenland and the icecaps of Ellesmere Island, Iceland, Spitsbergen, and Novaya Zemlya, with almost no sizable icebergs produced along the Eurasian or Alaskan Arctic coasts. No icebergs are discharged or drift into the North Pacific Ocean or its adjacent seas, except for a few small ones that calve each year from the piedmont glaciers along the Gulf of Alaska. These achieve no significant drift. *See* ARCTIC OCEAN; ATLANTIC OCEAN.

The Arctic ocean currents and adjacent seas determine the drift and ultimately, the distribution of icebergs—wind having only short-term effects. Icebergs along the East Greenland coast drift predominantly southward around the tip of Greenland (Cape Farewell) and then northward along its west coast. Here they join the West Greenland icebergs and travel in a counterclockwise gyre across Davis Strait and Baffin Bay. The icebergs are then carried southward under the influence of the Labrador Current along the coasts of Baffin Island, Labrador, and Newfoundland. This drift terminates along the Grand Banks of Newfoundland, where the waters of the Labrador Current mix with the warm northeastward flowing Gulf Stream. Here even the largest icebergs melt within 2–3 weeks. *See* GULF STREAM.

Freak iceberg drifts, outside the limits shown in Fig. 1, have been reported in the mid-Atlantic and off Scotland, Nova Scotia, and the Azores. In 1926 the southernmost known iceberg (a growler) reached 30–20°N, 62–32°W (about 150 mi from Bermuda). In 1912 a growler was sighted 75 mi east of Chesapeake Bay, U.S.A. Such sightings, however, are extremely rare.

About 500 icebergs a year are carried past Newfoundland (latitude 48°N) and into the North Atlantic, usually between February and June. These are survivors of an estimated 3-year journey from West Greenland. Most icebergs become stranded along the Arctic coasts and are ultimately destroyed by wave action and summer melting. The number arriving at the Grand Banks each year varies from near zero (in 1924, 1940, 1941, 1958) to 1500 or more (in 1972, 1984, 1991, 1994).

Icebergs in the Northern Hemisphere may reach proportions as large as 2000 ft in width and 400 ft in height above the water (**Fig. 2**). Such icebergs weigh upward of 15 million tons. More commonly, icebergs are 200–500 ft in width and 200 ft in height. The highest iceberg ever sighted (1959, in Melville Bay, Greenland) was 550 ft above the water. Shipping hazards, however, come not so much from large icebergs as from smaller ones, which can be difficult to detect by radar.

Ice islands. An icecap extending seaward on a broad front may form a landfast floating ice shelf rather than as an iceberg-producing glacier. Such ice shelves are common in the Antarctic but relatively rare in the Arctic. Ice broken from the shelf drifts freely as large tabular masses, ranging in size

Fig. 2. Arctic iceberg, eroded to form a valley or drydock type; grotesque shapes are common to the glacially produced icebergs of the North. Note the brash and small floes of sea ice surrounding the berg.

from a few square miles to several hundred square miles (**Fig. 3**). The term “ice island” is limited to the Arctic. Similar masses in the Antarctic are termed icebergs. Ice islands originate from the ice shelves of the northern shores of the Canadian archipelago and Greenland, and usually circulate in the Arctic Ocean for an indefinite time. The principal source of large ice islands is the Ward Hunt Ice Shelf on Northern Ellesmere Island. Aerial reconnaissance since 1946 has observed about 100 ice islands. They occasionally drift into the Norwegian Sea and Baffin Bay, where they erode and become icebergs. During

Fig. 3. Antarctic iceberg, tabular type. Such bergs develop from great ice shelves along Antarctica and may reach over 100 mi (160 km) in length. The U.S. Coast Guard icebreaker *Westwind* is in the foreground.

World War II, Winston Churchill proposed establishing airfields on ice islands in the Norwegian Sea. This was not done because of the faster rate of melting in those waters. Scientific camps have been set up on ice islands mostly by United States and Russian research agencies.

International Ice Patrol. The sinking of the passenger liner *Titanic* in 1912 with a loss of over 1500 lives resulted in the formation of the International Conferences on the Safety of Life at Sea. This included the establishment of the International Ice Patrol. The Patrol, operated by the U.S. Coast Guard, maintains an ice observation and reporting service in the North Atlantic Ocean. It is supported by 20 nations and has operated each year (except wartime) since 1913. First by ships, then search aircraft, the Patrol now uses satellite imagery and airborne radar as observation tools. It also has carried out oceanographic surveys, Greenland glacier mapping, and iceberg destruction experiments.

Other ice reporting services are maintained by Canada, Denmark, Iceland, Finland, Japan, and Russia for their respective waters, and the U.S. National Ice Center for Polar Regions.

Antarctic icebergs. In the Southern Ocean, icebergs originate from the giant ice shelves along the Antarctic continent and from glaciers which may flow into the shelves or form floating ice tongues. Masses breaking from either can result in huge, tabular icebergs 200 ft high and several hundred square miles in area. The world's largest glacier and iceberg producer is the Lambert Glacier (70°E longitude), which merges with the Amery Ice Shelf. In 1963 a giant iceberg measuring 50 by 70 mi was calved. This iceberg was tracked until 1976, by which time it had broken into thousands of smaller icebergs. The largest iceberg ever reported was calved in 1992 from the Thwaites Ice Tongue (110°W longitude). It was larger than the state of Rhode Island, and was still being tracked in 2000 with a size of about 25 by 50 mi.

Antarctic icebergs drift throughout the Southern Ocean under the influence of the Antarctic Circumpolar Current, and into the South Atlantic, Indian, and Pacific oceans, where they have been sighted as far north as latitude 35°S (Fig. 1). When weathered, Antarctic icebergs attain a deep bluish hue of great beauty rarely seen in the Arctic.

Water volume. The total volume of ice calved each year as icebergs is estimated to equal about half the volume of the world's total water usage. The feasibility of towing or otherwise moving large icebergs to water-deficient lands has been proposed and studied. But the difficulties of towing large icebergs over 6 million tons (as determined by actual tests), and the rapid deterioration of icebergs in warm waters appear to make such a venture impractible at this time. Robertson P. Dinsmore

Bibliography. N. Bowditch, Ice in the sea, The American Practical Navigator, chap. 34, *U.S. Defence Mapping Agency Publ.*, no. 9, 1995; T. F. Budinger, Iceberg detection by radar, *Proceedings of the International Maritime Consultative Organization Conference on Safety of Life at Sea*, London, 1960; R. P. Dinsmore, Ice and Its Drift into the North Atlantic Ocean, *International Commission for Northwest Atlantic Fisheries (ICNAF) Spec. Publ.*, no. 8, 1972; J. Sater, Arctic Drifting Stations, *Report on Activities Supported by the Office of Naval Research*, 1968; E. H. Smith, The Marion Expedition to Davis Strait and Baffin Bay, 1928-29-30-31, 3 vols., *U.S. Coast Guard Bull.*, no. 19., 1932; U.S. Coast Guard, Annual Reports of the International Ice Patrol, *USCG Publ.*, no. 188, 1919 onward; U.S. Defense Mapping Agency, Sailing Directions for Antarctica, *DMA Publ.*, no. 200, 1992; World Meteorological Organization, Sea Ice Nomenclature, *WMO Publ.*, no. 259, 1971.

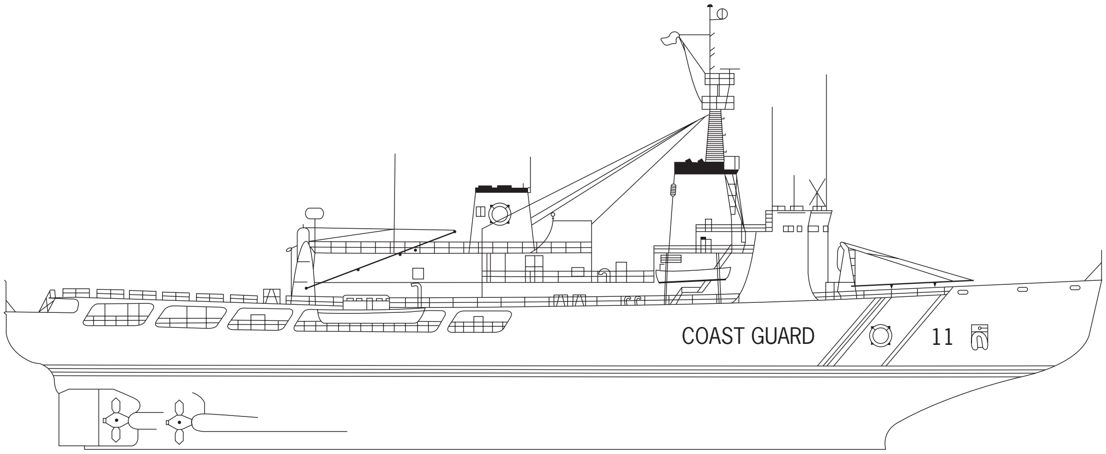
Icebreaker

A ship designed to break floating ice. Icebreaker technology has evolved rapidly since the 1960s as a result of potential resource development in the Arctic regions of Russia, Canada, and the United States. This led to the construction of icebreaking ships that can transit to all areas of the world, including the North Pole. The Arctic Ocean and Antarctica play a large role in shaping the global climate. With the Arctic Ocean and the ice-covered waters of Antarctica being the least studied of all the oceans, icebreakers provide the platforms from which polar science and research can be conducted on a year-round basis. On August 22, 1994, the first U. S. and Canadian icebreakers, USCGC *Polar Sea* and CCGS *Louis S. St. Laurent*, arrived at the North Pole after having conducted scientific research in areas never previously studied. *See* ANTARCTIC OCEAN; ARCTIC OCEAN.

Types. Icebreakers may be classed as polar or subpolar, depending on their primary geographic area of operation. Polar icebreakers have a capability for independent operation in first-year, second-year, and multiyear ice in the Arctic or the Antarctic. Subpolar icebreakers are capable of icebreaker operations in the ice-covered waters of coastal seas and lakes outside the polar regions. *See* MARITIME METEOROLOGY; SEA ICE.

Resistance. The speed that an icebreaker can achieve in ice depends upon the resistance of the hull to break the ice and submerge the broken ice pieces. The resistance increases approximately as the square of the level ice thickness. As an example, the resistance in breaking 4 ft (1 m) of ice is basically four times that for breaking 2 ft (0.5 m) of ice. Resistance increases almost linearly with an increase in the ship's beam. Other factors that influence resistance are the amount of friction developed between the hull surface and the ice, and, to a lesser degree, the length of the icebreaker.

Hull form. The combined shapes of the bow, midbody, and stern constitute the hull form. Each component has unique requirements to make it efficient in icebreaking and ice transiting. The icebreaker's bow must be designed to break ice efficiently, and



Outboard profile of the U.S. Coast Guard icebreaker *Polar Sea*.

it consists of an inclined wedge that forms an angle of about 20° with the ice surface (see **illus.**). As the icebreaker advances, the bow rides up on the edge of the ice until the weight of the ship becomes sufficiently large that the downward force causes the ice to fail. As the icebreaker advances, the broken ice pieces move under the hull and the process is repeated. The sides of the icebreaker can be vertical or sloped. Traditionally, icebreakers have sloped sides of 5–20° to reduce the frictional resistance. Sloped sides can also improve the turning ability (maneuverability) of the ship in ice. The stern of the icebreaker is designed in a manner similar to the bow so that the ship can break ice while going astern. It must also be designed for the proper flow of water into the propellers in both the ahead and astern directions. Because of the many factors that influence icebreaking performance, the design involves considerable compromise and judgment, particularly when icebreakers may spend 50% of their time in open water. *See* SHIP DESIGN.

Hull structure. The hull of a polar icebreaker is built very strong so that it can withstand tremendous ice forces. Many icebreakers have to withstand impacts with ice that is 20 ft (6 m) thick. Because these frequent and high loads occur at very low air temperatures of −50°F (−45°C), specialty steels are used on polar icebreakers. It is common to have bow and stern plating of 1.5–2.0 in. (4–5 cm) in thickness. The side and bottom plating thickness are somewhat less. Subpolar icebreakers, with only a limited ice transiting capability, have only a portion of the hull strengthened. It is called an ice belt and runs several feet above and below the waterline of the hull.

Propulsion. Icebreaker propulsion systems consist of engines, transmission systems, and propellers. Most icebreakers are powered with diesel engines. However, when the horsepower required for icebreaking is sufficiently great, gas turbines or nuclear steam turbines are needed. Transmission systems are either mechanical (primarily gears) or electrical (generators and motors). Propellers on an icebreaker can number from one to four, and the propellers can be either fixed pitch or controllable pitch. A single propeller is frequently used on those ships requiring only limited icebreaking capability. Most icebreakers, however, have either two or three propellers for purposes of system reliability and improved maneuverability. All propulsion systems must be capable of withstanding the shock and high torque loads that are transmitted to the system when the propeller blades impact or mill ice pieces. *See* MARINE ENGINE; MARINE MACHINERY; PROPELLER (MARINE CRAFT); SHIP POWERING, MANEUVERING, AND SEAKEEPING.

Reduction of icebreaking resistance. Many auxiliary methods have been developed to improve the performance of icebreakers by reducing the friction between the hull and the ice. One widely used method is to roll or heel the ship from side to side. The rolling motion is achieved by pumping large quantities of water from side to side in specially configured heeling tanks. In addition to heeling systems, most icebreakers use a low-friction coating or a stainless-steel ice belt to reduce icebreaking resistance.

Another concept is the use of air bubbler systems. In these systems, icebreaking resistance is reduced by discharging large volumes of air underwater at the bow and sides of the icebreaker. The rising air bubbles carry large quantities of water to the surface and reduce the friction at the hull-ice interface. Another method to reduce friction is the water wash system, involving pumping large quantities of water through nozzles located in the hull onto the ice surface at the bow and forward sides of the icebreaker.

Tests in ice model basins. The ability to evaluate the icebreaking performance of new designs before construction is very important. A scaled model of the ship is made and tested in a towing basin inside a large refrigerated space. Ice is grown on the water surface, and the model is towed or self-propelled through the ice to determine the resistance. *See* TOWING TANK.

Richard P. Voelker

Bibliography. L. W. Brigham (ed.), *The Soviet Maritime Arctic*, 1991; R. A. Dick and J. E. Laframboise, An empirical review of the design and performance of icebreakers, *Mar. Technol.*, 26(2):145–159, 1989; E. V. Lewis (ed.), *Principles of Naval*

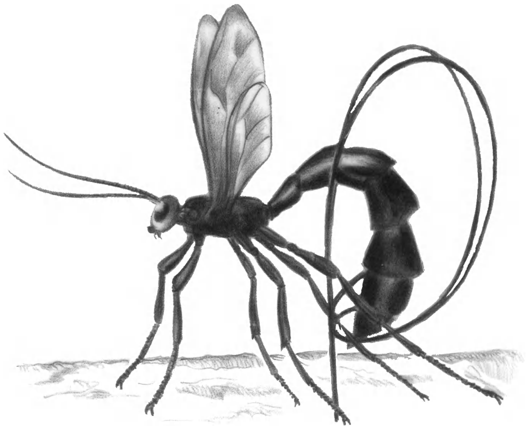
Architecture, 2d ed., 3 vols., Society of Naval Architects and Marine Engineers, 1988; R. P. Voelker et al., Eight years of performance measurements aboard *Polar* class icebreakers, *Soc. Nav. Arch. Mar. Eng. Trans.*, 100:271–292, 1992; H. Yamaguchi, Experimental Voyage through Northern Sea Route, INSROP Symposium Tokyo '95, 1996.

Ichneumon

The common name for medium- to large-sized insects (wasps) comprising 25,000 species and belonging to the family Ichneumonidae (Hymenoptera); together with the Braconidae (with an additional 15,000 species) they form the superfamily Ichneumonoidea. The majority of ichneumons are slender wasps with long filiform antennae, freely rotating head, and permanantly extruded ovipositor. The ichneumons, found worldwide, are most abundant in heavy vegetational areas such as the edges of woodlands, hedgerows, and in secondary growth vegetation, and are often seen in the woodland canopy or in the litter of the forest floor. Ichneumons are more commonly seen in the spring and fall, as these wasps avoid hot, dry weather.

All members of the family Ichneumonidae are parasitic during their immature stages. They attack a wide variety of insects, usually the larval stages of beetles, butterflies, moths, flies, and other Hymenoptera. Since many of their insect hosts are considered pests of forest and agricultural crops, ichneumons are beneficial and important in biological control. *See* ENTOMOLOGY, ECONOMIC.

Female ichneumons use chemical and physical cues emanating from the host (the insect on or within which the ichneumon's progeny will develop) to locate it for egg laying. The length of the ovipositor is variable, being relatively short in species attacking exposed hosts, and extremely long in species attacking hosts enclosed in galls or stem cavities, or boring in wood. Some species, such as in the genus *Megarhyssa* (see **illus.**), may have ovipositors longer than the body.



Long-tailed ichneumon (*Megarhyssa lunator*), about 6 in. (15 cm), of which more than 4 in. (10 cm) is "tail."

The female ichneumon deposits an egg on or within the host. Most species are solitary; that is, only one egg is laid. A few species are gregarious as larvae, with the female laying several eggs in the host which all develop into adults. Species that lay their eggs on the host are referred to as ectoparasites; those which deposit their eggs within the host, endoparasites. The eggs hatch into white, fleshy, grublike larvae which slowly consume the host. The ichneumon larvae feed through the cuticle if they are ectoparasites or feed internally on tissues if they are endoparasites. The mature larvae usually kill the host as they emerge and subsequently pupate in a cocoon.

The females of some species of ectoparasitic ichneumons inject a venom that paralyzes the host prior to laying an egg. Some endoparasites have been found to inject a symbiotic virus along with an egg into the host. The virus appears to alter the physiology of the host for the parasite's benefit.

Most ichneumons are host-specific; that is, they only attack and develop in a particular species of host. However, a few species are polyphagous, and attack a number of different hosts. The ability to attack and destroy a wide variety of insects makes ichneumons particularly important to humans in suppressing pest insects. *See* HYMENOPTERA; INSECTA.

S. B. Vinson

Bibliography. R. P. Askew, *Parasitic Insects*, 1971.

Ichthyopterygia

A supraorder of extinct marine reptiles from the Mesozoic, or the Age of Dinosaurs, belonging to the subclass Diapsida. The group consists of the order Ichthyosauria and several basal forms from the Early Triassic. *See* DIAPSIDA; DINOSAUR; REPTILIA.

Fossil record. Ichthyopterygians existed for about 160 million years. Their oldest definitive record comes from the latest part of the Early Triassic (latest Olenekian), whereas the youngest record is from the strata formed during the earliest part of the Late Cretaceous (Cenomanian). Ichthyosaurs became extinct about 25 million years before dinosaurs disappeared at the end of the Cretaceous. The timing approximately corresponds to an extinction event of marine organisms at the Cenomanian/Turonian boundary. *See* EXTINCTION (BIOLOGY).

Body characteristics. Ichthyopterygians are noted for their tuna-shaped (tunniform) body outline in combination with the presence of a well-defined tail fin or caudal fluke. Cetaceans (whales and dolphins) were the only other tetrapods (vertebrates with four limbs) to evolve this body form. Early ichthyosaurs were elongate and could be loosely described as "lizards with flippers." These early forms probably propelled themselves using body undulation, as in living cat sharks. Various body outlines evolved during the Middle Triassic, but only a single lineage, which gave rise to a demarcated caudal fluke, survived the Late Triassic. A tuna-shaped body outline also became established by mid Late Triassic time (middle Norian), and this group diversified into a large number of species during the Jurassic and Cretaceous.

Body length varied from approximately 0.5 m (2 ft) to over 20 m (67 ft) in ichthyosaurs. Body size of the first ichthyosaurs (Early Triassic) never exceeded 3 m (10 ft), but giants reaching 15 m (50 ft) or more appeared during the Late Triassic (*Shonisaurus*), and some fragments suggest the possibility of a second whale-sized form in the Early Jurassic. The largest species from the Jurassic and Cretaceous attained maximum lengths of 10 m (33 ft).

Eyes. Many Jurassic ichthyosaurs possessed disproportionately large eyeballs, including the largest eyes that have ever been measured in vertebrates, which reached over 26 cm (10 in.) in diameter in *Temnodontosaurus*, whose body length was about 9 m (30 ft). *Ophthalmosaurus* had eyes measuring 23 cm (9 in.) and had a body length of less than 4 m (13 ft). *See* EYE (VERTEBRATE).

Limbs. Early ichthyosaurs had five digits (fingers) in their hands and feet (that is, their front and back flippers), as in most other tetrapods. However, some later species had as many as 10 digits per hand, whereas others had as few as three. These ichthyosaurs also had disk-shaped phalanges and metacarpals (finger and palm bones), unlike most other vertebrates with limbs.

Diet. Ichthyopterygian diet included fish, cephalopods, shells, and some other vertebrates such as sea turtles and birds. Possibilities also existed for feeding upon other marine reptiles. R. Motani

Bibliography. C. McGowan and R. Motani, Ichthyopterygia, *Handbuch der Palaerpetologie Part 8*, Verlag Dr. Friedrich Pfeil, Munich, 2003; C. McGowan, *Dinosaurs, Spitfires, and Sea Dragons*, Harvard University Press, Cambridge, MA, 1991; R. Motani, Ichthyosauria: Evolution and physical constraints of fish-shaped reptiles, *Annu. Rev. Earth Planet. Sci.*, 33:12.1–12.26, 2005; R. Motani, Rulers of the Jurassic seas, *Sci. Amer.*, 283(6):52–59, December 2000.

Ichthyornithiformes

A group of extinct flying birds known exclusively from the latest stages of the Cretaceous (80–65 million years ago) in North America. Along with Hesperornithiformes, Ichthyornithiforms are some of the earliest nonmodern bird taxa known to science. Indeed, alongside *Archaeopteryx* (from much earlier Jurassic rocks in Germany, about 140 million years old), the ichthyornithiformes *Ichthyornis*, *Apatornis*, and their kin formed the basis for knowledge of avian evolution for almost a century. Although the fossil record of these birds has improved in recent years, Ichthyornithiformes are still poorly understood. *See* AVES; HESPERORNITHIFORMES; ODONTOGNATHAE.

The discovery of *Ichthyornis* in 1875 caused excitement—here was a flying bird that possessed sharp pointed teeth. Charles Darwin wrote about Ichthyornithiformes (and their relatives, Hesperornithiformes) in 1880 and noted that they appeared to provide excellent evidence in support of his ideas on the origin of species. Most recently, these archaic avians were placed within the clade Ornithurae, which also contains Hesperornithiformes and all modern birds (Neornithes). *See* NEORNITHES.

Much controversy has surrounded these birds: only recently have they been well described anatomically, and at least one familiar skeleton appears to comprise several individuals. Although all original finds of these birds were from the Cretaceous marine rocks of Kansas, more recent discoveries have extended their known range across the Northern Hemisphere, into Europe and Asia. Ichthyornithiformes appear to have been important and diverse constituents of late Cretaceous ecosystems, dying out along with most lineages of dinosaurs at the end-Cretaceous mass extinction event. *See* EXTINCTION (BIOLOGY). Gareth Dyke

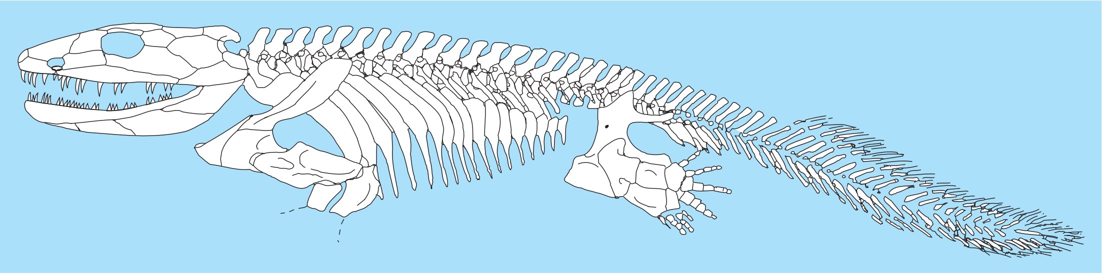
Bibliography. S. Chatterjee, *The Rise of Birds*, Johns Hopkins University Press, Baltimore, 1997; L. M. Chiappe and L. M. Witmer, *Mesozoic Birds: Above the Heads of Dinosaurs*, University of California Press, Berkeley, 2002; A. Feduccia, *The Origin and Evolution of Birds*, 2d ed., Yale University Press, New Haven, 1999; S. L. Olson, *The Fossil Record of Birds*, in *Avian Biology*, vol. 8, Academic Press, New York, 1985.

Ichthyostega

Four-legged vertebrates (basal tetrapods) that evolved from their lobe-finned fish ancestors during the later Devonian Period (400–350 million years ago) [see **illus.**]. *Ichthyostega* was the first Devonian tetrapod to be described and was found in East Greenland during the 1930s. For many decades it remained the only representative of the “fish-tetrapod transition” which was known from articulated skeletal fossils. Subsequently, a second genus from the same beds, *Acanthostega*, was more fully described, and during the 1990s Devonian tetrapod genera were recognized in other parts of the world.

Like other very early tetrapods, *Ichthyostega* had a skull composed of an outer casing of bone (the skull roof) with a separate box containing the brain (the braincase) inside it. The skull roof bore the teeth, and the teeth of the upper jaw were substantially larger but fewer in number than those of the lower jaw. This suggests some specialization of diet, but it is not clear for what purpose. *Ichthyostega* was undoubtedly a predator. The braincase shows some highly unusual features, especially in the ear region, which are difficult to interpret and do not resemble those of any other early tetrapod.

The postcranial skeleton shows some primitive and some very specialized features. The tail bore a fringe of fin-rays like those of a fish; these are lost in all other known early tetrapods except *Acanthostega*. The vertebrae consisted of several parts. The centra were formed of hoops of bone surrounding the



Reconstruction of the skeleton of *Ichthyostega* based on recent information. The animal was about 1 m (3 ft) long.

flexible supporting notochord. The notochord was the main point of attachment between the vertebral column and the head, inserting into the rear of the braincase. The neural arches supporting the spinal cord articulated with one another via well-developed joint surfaces (zygapophyses). The ribs were massive overlapping blades which formed a corset around the body and gave it rigidity. The animal would not have been able to bend very much.

The massive shoulder girdle bore the large complex humerus which articulated with the shoulder at right angles to the body. The elbow was also more or less permanently fixed in a right angle, allowing the forearm to be used as a prop but permitting very little other movement. The hand is still unknown. The large pelvic girdle was attached to the vertebral column only by ligaments and muscles—no bony joint surfaces were involved as they are in modern tetrapods. The femur appears to have been about half the length of the humerus, and the hindlimb was paddlelike with tibia, fibula, and ankle bones all broad and flattened. The foot bore seven toes, with three tiny ones at the leading edge and four stout ones behind. It seems likely that the toes were enclosed in a web of skin.

Because of its peculiarities, *Ichthyostega* has not been helpful in working out many ideas about the origin of tetrapods. Now however, it can be seen in the context of other Devonian tetrapods, and its specializations will help provide more information about how, where, and when early tetrapods first radiated onto the land. *See* AMPHIBIA. Jennifer A. Clack

Bibliography. J. A. Clack, Devonian tetrapod trackways and trackmakers: A review of the fossils and footprints, *Palaeogeog. Palaeoclimatol. Palaeoecol.*, 130:227–250, 1997; E. Jarvik, The Devonian tetrapod *Ichthyostega, Fossils and Strata*, 40:1–213, 1996.

Igneous rocks

Rocks that are formed from melted rock, usually by cooling. The melted rock is called magma when it is below the surface, and lava when it flows out onto the surface. When the magma or lava cools or moves into areas of lower pressure, crystals or glass or both form. Thus, igneous rocks are composed of crystals, glass, or combinations of these materials. Magmas sometimes erupt explosively, creating ash that is composed of broken crystals, glass, and rock materials called pyroclastic material. Rocks formed at or very near the surface, including pyroclastic rocks, are called volcanic rocks, whereas those formed from magma at depth are called plutonic rocks. In the past, some rocks that formed below the surface but near it were called hypabyssal rocks. *See* LAVA; MAGMA.

Like all rocks, igneous rocks are classified on the basis of their composition and texture. Both mineral and chemical compositions are used for classification, but the rocks are first subdivided by texture into three categories. (1) Volcanic rocks are dominated by interlocking grains too small to see with the unaided eye or a simple lens. (2) Plutonic rocks are composed almost entirely of interlocking grains big enough to see. (3) Pyroclastic rocks are composed of fragments (clasts) of glass, minerals, and preexisting rocks. *See* PETROLOGY.

Composition. The range of igneous rock chemistries is relatively large. The common igneous rocks consist of silica (SiO_2), 40 to 77%; alumina (Al_2O_3), <1% to nearly 20%; magnesia (MgO), <1 to 50%; total iron oxide, <1% to more than 10%; and lime (CaO), soda (Na_2O), and potash (K_2O), <1% to 9% (see **table**).

The minerals in common igneous rocks include quartz, alkali and plagioclase feldspars, micas (muscovite, biotite, and phlogopite), amphiboles (such as hornblende and riebeckite), pyroxenes (including augite and the orthopyroxenes), and olivines. In addition, minerals occurring in small amounts include apatite, zircon, epidote, titanite, magnetite, and ilmenite. *See* AMPHIBOLE; FELDSPAR; MICA; OLIVINE; PYROXENE; QUARTZ.

For purposes of description, igneous minerals are divided into three groups. Essential minerals are those on which the rock name is based. Characterizing accessory minerals are nonessential minerals that occur in amounts of 5% or more. Minor accessory minerals occur in amounts of less than 5%. *See* MINERAL.

Classification. Traditionally, classification of all igneous rocks was based on minerals and textures. In volcanic rocks, however, the minerals are often so small that identification is difficult or impossible, and glass contains no minerals. Thus, in recent years volcanic rock classification has been based largely on rock chemistry, except for classifications used in field studies.

Volcanic flow and pyroclastic rocks. The International Union of Geological Sciences (IUGS) volcanic rock

Chemical compositions of some igneous rocks in wt %

Compound	Dunite[a]	Basalt[b]	Andesite[a]	Foid-syenite[c]	Granite[a]	Rhyolite[d]
SiO_2	40.08	49.8	58.97	59.54	69.22	77.24
TiO_2	0.01	2.6	1.04	0.14	0.48	0.20
Al_2O_3	0.29	14.0	17.17	18.60	15.50	10.81
Fe_2O_3	0.31	2.5	4.36	2.86	1.03	1.66
FeO	7.62	8.5	2.02	2.09	1.42	0.27
MgO	49.69	7.2	1.51	0.10	0.73	0.33
CaO	0.11	11.3	4.90	1.16	1.93	1.48
Na_2O	0.05	2.2	4.23	8.96	4.15	2.59
K_2O	0.01	0.62	2.90	4.24	4.42	4.12
P_2O_5	0.00	0.32	0.51	0.16	0.15	0.06
Other	0.58	0.35	1.55	1.80	0.30	0.65
TOTAL	98.86	99.6	99.26	99.87	99.37	99.80

[a] F. J. Flanagan (ed.), Descriptions and Analyses of Eight New USGS Rock Standards, *U.S. Geol. Surv. Prof. Pap.*, no. 840, 1976.
[b] F. J. Flanagan et al., Basalt, BHVO-I, from Kilauea Crater Hawaii, in F. J. Flanagan (ed.), *U.S. Geol. Surv. Prof. Pap.*, no. 840, 1976.
[c] P. D. Snavely, Jr., et al., Nepheline syenite, STM-1, from Table Mountain, Oregon, in F. J. Flanagan (ed.), *U.S. Geol. Surv. Prof. Pap.*, no. 840, 1976.
[d] M. H. Staatz and W. J. Carr, Geology and Mineral Deposits of the Thomas and Dongway Ranges, Juab and Tooele Counties, Utah, *U.S. Geol. Surv. Prof. Pap.*, no. 415, 1964.
SOURCE: From L. A. Raymond, *Petrology: The Study of Igneous, Sedimentary, and Metamorphic Rocks*, 2d ed., 2002.

classification defines the common volcanic rock types on the basis of silica (SiO_2) content versus alkalis ($Na_2O + K_2O$). The most common types of volcanic rocks have less than 7% total alkalis and include basalt (45–52% SiO_2), basaltic andesite (52–57% SiO_2), andesite (57–63% SiO_2), dacite (63% to 69–77% SiO_2), and rhyolite (>69% SiO_2). *See* ANDESITE; BASALT; DACITE; RHYOLITE.

Pyroclastic volcanic rocks are classified on the basis of grain size, in addition to the compositional name. If the dominant grain size is less than 2 mm, the volcanic material is called ash and the rock composed of these grains is called tuff. If the rock contains a significant percentage of grains between 2 and 64 mm, it is called a lapilli tuff. Rocks dominated by fragments larger than 64 mm are called volcanic breccias. Thus, a complete pyroclastic rock name might be lapilli tuff rhyolite. *See* BRECCIA; PYROCLASTIC ROCKS; TUFF.

Some volcanic rocks are composed entirely or almost entirely of glass. If the glass largely lacks vesicles (gas bubble holes), the rock is called obsidian (or rhyolite obsidian). If the glassy rock contains abundant, small, aligned, tubular holes, it is called pumice. Larger vesicles, characteristic of the less siliceous rocks, require the modifier "vesicular," as in vesicular basalt. Rocks dominated by vesicles are called (basalt) scoria. *See* OBSIDIAN; PUMICE.

Plutonic rock. Classification of plutonic rocks is based primarily on mineralogy. The IUGS classification is the most widely used plutonic rock classification. In this classification, which features two triangles joined at their bases, the essential minerals are quartz, alkali feldspar, plagioclase, and feldspathoid minerals (foids). All other minerals are ignored for classification purposes, except in classification of low-silica rocks, for which separate sets of triangles are used. *See* FELDSPATHOID.

In order to classify a rock, the essential mineral percentages are observed and totaled, and the individual percentages are divided by the total to obtain a percentage of the three essential minerals. The single point on the triangle represented by the three mineral percentages determines the name for the rock. Among the most common plutonic rocks are granite, composed of quartz, alkali feldspar, and some plagioclase feldspar; and gabbro, composed of augite and plagioclase (**Fig. 1**). In the United States, some geologists use alternative classifications that subdivide the "granite" field of the IUGS classification. *See* GABBRO; GRANITE.

For low-silica rocks—the gabbros and ultrabasic rocks—separate triangles are used for classification (**Fig. 2**). The essential minerals for these classifications are plagioclase, olivines, pyroxenes, and hornblende. *See* HORNBLENDE.

Plutonic rocks are also classified by some geologists on the basis of their chemistry. Thus, there are peraluminous, metaluminous, and peralkaline rocks, which have chemistries high, moderate, and low in

Plutonic rocks	GRANITE	QUARTZ DIORITE	DIORITE	GABBRO	PERIDOTITE
Volcanic rocks	Rhyolite	Dacite	Andesite	Basalt	Komatüte

Fig. 1. Some plutonic rock names and their volcanic equivalents.

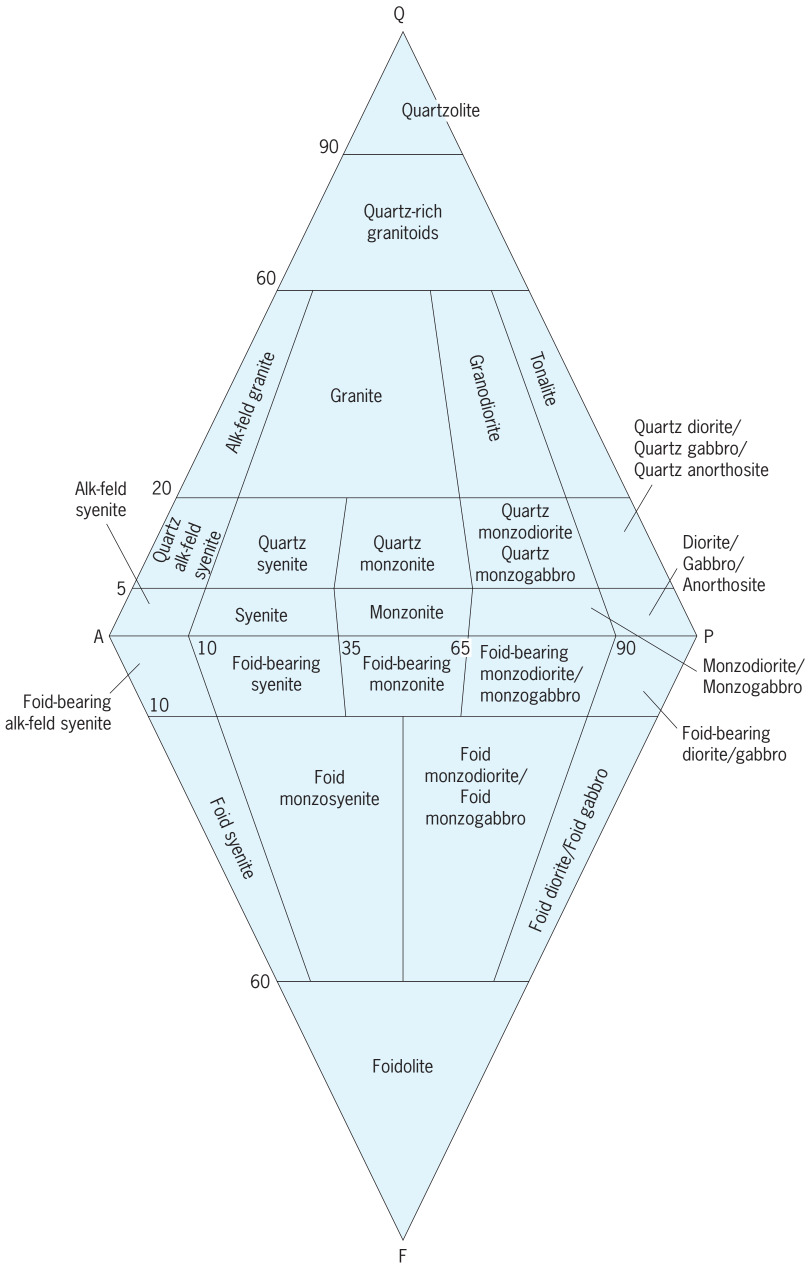


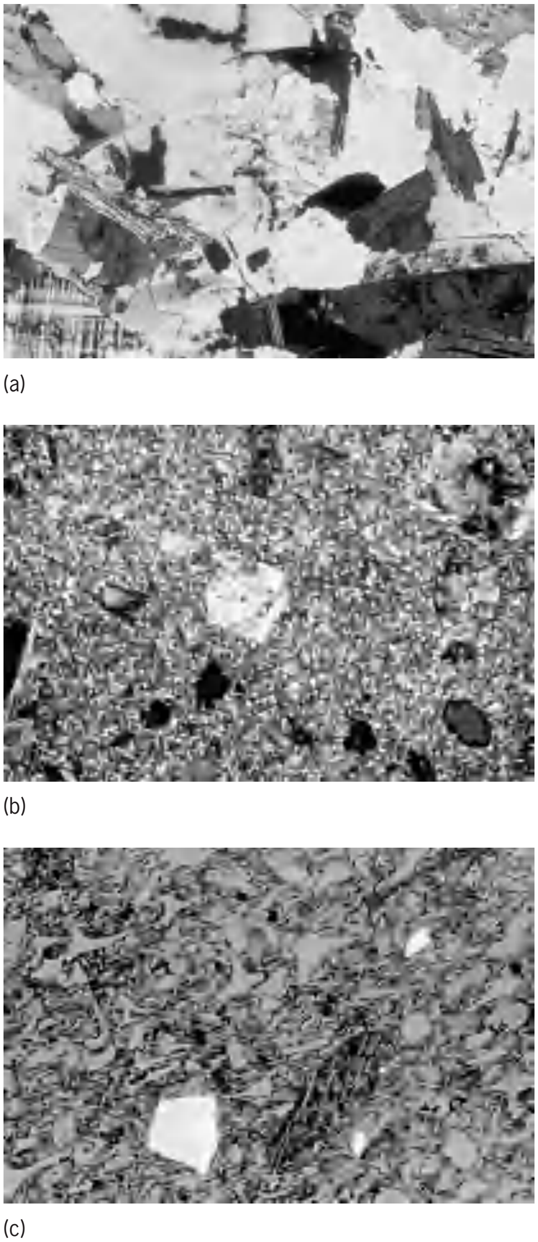
Fig. 2. IUGS classifications of gabbroic rocks (triangle AQP) and ultramafic rocks (triangle AFP). (*From A. Streckeisen, To each plutonic rock its proper name, Earth Sci. Rev., 12:1–33, 1976*)

alumina, respectively. And there are S-, I-, and A-type granitoid rocks formed from melts of sedimentary, igneous, and relatively high-alkali materials.

Textures and structures. The texture of a rock includes the size, shape, arrangement, and grain boundary features. Many geologists also include abundant holes and the relationship between glass and crystals in volcanic rocks as part of the texture. In general, a feature is textural if there are many of them in a hand-sized sample. Structures are generally larger features. There may be a few or only one in a hand specimen, or they may be much larger than a hand specimen. *See* PETROGRAPHY.

All of the plutonic rocks are crystalline in texture. This means that the grain boundaries in the rock are interlocking (**Fig.** 3*a*). Rocks with grains large enough to see are said to have phaneritic texture. If the grain length exceeds 3 cm in length, the rock has pegmatitic texture. Volcanic rocks include those with crystalline textures too small to see without magnification (aphanitic textures), those with aphanitic crystals surrounding larger visible crystals (porphyritic textures) [Fig. 3*b*] and those composed of fragments (clasts) formed by explosive eruption of magmas (pyroclastic textures) [Fig. 3*c*]. In the last, clasts may include fragments of glass, fragments of crystals, or fragments of earlier-formed rocks.

Grain size, shape, and arrangement. Grain (crystal) sizes in igneous rocks vary widely. The smallest crystals


(a)

(b)

(c)

Fig. 3. Photomicrographs of common igneous rock textures. (*a*) Hypidiomorphic-granular texture. (*b*) Porphyritic texture. (*c*) Pyroclastic texture. Width of field of view is 3.2 mm (0.125 in.).

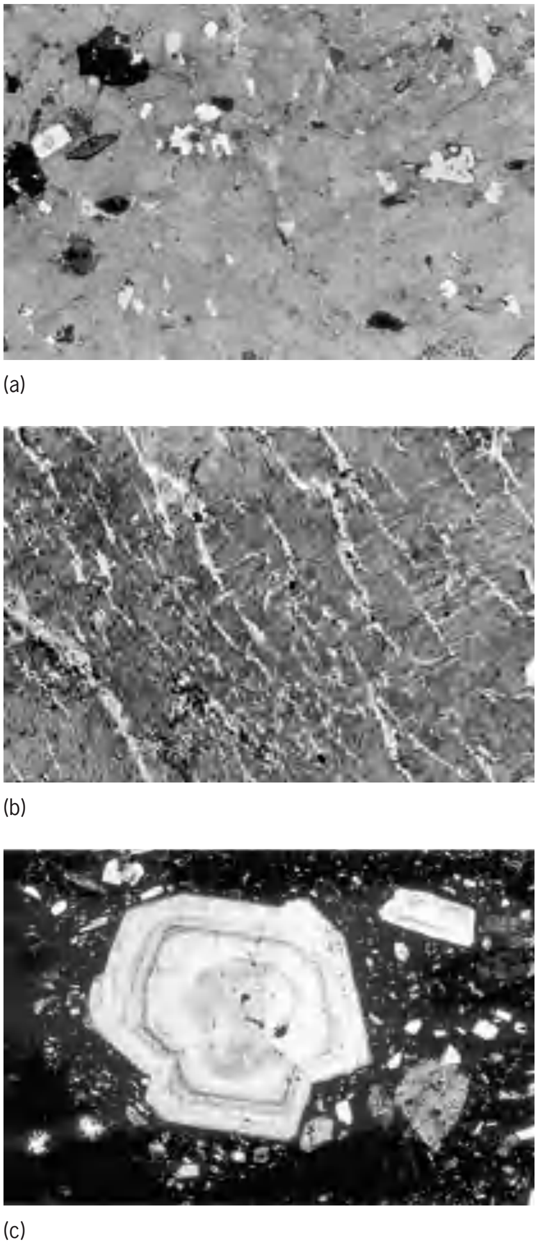


Fig. 4. Photomicrographs of some mineral textures in igneous rocks. (*a*) Poikilitic texture in alkali feldspar. (*b*) Perthitic texture in alkali feldspar. (*c*) Zoned texture in plagioclase. Width of field of view is 3.2 mm (0.125 in.).

are less than 0.01 mm, whereas the largest crystals are meters long. These largest crystals form in pegmatitic rocks, especially those rich in SiO_2. Fine-grained rocks have grains less than 1 mm in length. In medium-grained rocks, the average grains are 1 to 5 mm in length. Coarse-grained rocks have average grain lengths of 5 mm to 3 cm. In pegmatitic rocks, the dominant grain size exceeds 3 cm. In any given rock, there may be a range of crystal sizes for each mineral, called a crystal size distribution (CSD), which reflects the crystallization history. *See* PEGMATITE.

The crystalline grains of igneous rocks have shapes that vary and depend on a variety of factors controlling the growth history of the crystal. In addition, the crystal form of each mineral (the natural shape of the particular mineral) is a dominant influence on the shape. Mineral grains that have flat surfaces (crystal faces) on all sides, which formed as the crystal grew, are said to be euhedral. Crystals partially bounded by crystal faces are subhedral, and those not bounded by crystal faces are anhedral.

Early or rapid crystal growth from a melt yields partial, incomplete, or skeletal crystals. In volcanic glass, tiny beginning crystals of this type are called crystallites and microlites. In plutonic rocks, they are called skeletal or snowflake grains. *See* VOLCANIC GLASS.

Where growth of crystals is prevented by explosive eruption of magma, glass fragments erupt into the atmosphere. Tuffs form from accumulations of these variably sized, typically curved pieces of glass, mixed with fragments of crystals and rocks (Fig. 3*c*).

Minerals may form in random orientations, or they may have specific arrangements. In volcanic rocks in which elongate crystals are aligned (usually because of flow of the magma), the rock has a trachytic texture, but similar textures in plutonic rocks are called trachytoidal textures. Randomly oriented grains that are subhedral, on average, yield a hypidiomorphic-granular texture in plutonic rocks (Fig. 3*a*). This is the typical texture of the granites. The similar texture at a smaller scale in volcanic rocks is called intergranular texture.

Many specialized textures exist and have been named based on the kinds, shapes, or arrangement of grains. Porphyritic texture consists of two distinct sizes of grains, with substantially larger grains (phenocrysts) enclosed by a matrix of smaller grains (Fig. 3*b*). Porphyritic textures can be either volcanic or plutonic. Ophitic texture consists of large grains of pyroxene enclosing smaller, rectangular grains of plagioclase. Graphic texture, a texture that occurs within some pegmatitic rocks, consists of V- and L-shaped grains of quartz enclosed in a large alkali feldspar grain. The quartz in graphic texture is usually part of a single but incomplete skeletal grain. Corona textures are those in which a rim of one mineral surrounds a core of another mineral. *See* PHENOCRYST; PORPHYRY.

Mineral textures. Individual minerals within igneous rocks may also have special textures. Commonly in igneous rocks, larger grains of one mineral, such as alkali feldspar, will enclose many smaller grains of other minerals to form a poikilitic texture (**Fig. 4*a***). Ophitic and graphic textures are poikilitic textures that occur throughout and characterize particular rock types.

In granites, alkali feldspar grains also commonly have a mineral texture called perthitic. Perthitic texture results when small, tabular, irregular, or worm-shaped grains of plagioclase feldspar separate from, but typically occur as subparallel blobs within, larger alkali-feldspar grains (Fig. 4*b*). Myrmekitic and granophyric textures involve wormlike quartz intergrowths within plagioclase and alkali feldspar, respectively.


(a) (b) (c) (d)

Fig. 5. Some volcanic structures. (*a*) Lava plateau with flat-lying basalt lava flows, near Sprague, Washington. (*b*) Composite volcanoes, Atitlan (left) and Toliman in Guatemala. (*c*) Columnar jointing in andesite near Lost Creek Dam, Oregon. (*d*) Pillow basalts, Golden Gate National Recreation Area, California.

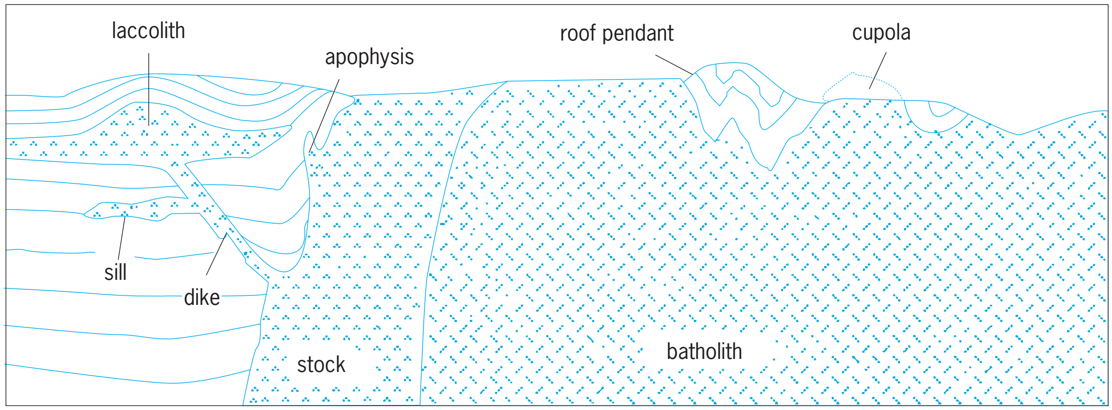


Fig. 6. Some plutonic structures.

Another important mineral texture is zoning (Fig. 4*c*). In this texture, common in plagioclase feldspar, each successive layer of the growing crystal has a different composition, so that the mineral appears to have rings of different colors or brightness under polarized light and magnification.

Texture formation. Various factors control the formation of crystals (mineral grains) in a magma. The temperature at which the crystal nucleates (begins to form), how fast it grows, the nature of the "neighborhood" of growth (that is, whether the crystal is surrounded entirely by melt or mostly by crystals, and whether the crystal is able to complete or nearly complete its growth) influence the shapes of crystals.

Crystal formation in a rock occurs in two stages: nucleation and growth. Some magmas retain unmelted crystal fragments of the rock from which the magma formed, and these form nuclei (clusters of the particular atoms needed to form specific minerals). When atoms simply attach themselves to preexisting surfaces or nuclei to form a growing nucleus for a mineral grain, this is called heterogeneous nucleation. If no nuclei exist in a magma, nuclei must form by establishing bonds between atoms. Nucleation from a magma containing no remnants of former crystals is called homogeneous. Heterogeneous nucleation is more common than homogeneous nucleation.

Once nuclei form, mineral grains grow by adding atoms to the nucleus. Growth is influenced by factors such as the composition of the melt, the temperature of the melt, the cooling rate during growth, and the rate of migration of atoms through the melt to the growing crystal. In high-silica magmas, crystals grow more slowly than in low-silica magmas, if all other factors are equal.

Some minerals begin to form earlier than others as a magma cools. For example, iron-and magnesium-rich minerals tend to nucleate at higher temperatures than potassium- and sodium-rich minerals. Once adequate cooling occurs, different minerals may grow simultaneously but at different rates. The final texture of a rock results from the growth of different crystals under the influence of the various controlling influences.

Volcanic structures. Volcanic and plutonic structures vary widely in size. Some are less than 1 cm in diameter, and others are thousands of meters across.

Major volcanic structures include lava plateaus and plains, pyroclastic sheets, shield volcanoes, composite volcanoes, and calderas (**Fig. 5**). Lava plains and plateaus are large sheets of lava that cover up to a million or more square kilometers. Pyroclastic sheets are similarly shaped masses of ash and cinder that have fallen from the sky as a result of explosive eruptions. Shield and composite volcanoes are cone-shaped piles of lava. Shield volcanoes are flat cones dominated by low-silica flows of basalt, whereas composite volcanoes are steep cones dominated by alternating lava flows and layers of pyroclastic material (Fig. 5*b*). Massive explosive eruptions sometimes produce large, crudely circular holes; these calderas may be tens to thousands of square kilometers in area. *See* CALDERA; VOLCANO; VOLCANOLOGY.

Intermediate-sized structures include small pyroclastic sheets, lava flows, pyroclastic (cinder) cones, and domes. Lava flows are sheets of lava that have flowed across the Earth's surface after erupting from a vent. Domes are piles of lava in the shape of an inverted cup that have not flowed far from the site of eruption because of their sticky character (high viscosity).

Small-sized structures range from tiny piles of lava (spatter cones) or cones of pyroclasts (small cinder cones) a few meters high to features less than a centimeter across that occur in hand specimens. Many of these structures are associated with lava flows. Small blobs of lava that squeeze out through cracks in a lava flow and solidify are called squeeze-ups. Larger bumps on lava flows are called pressure ridges. Where lavas cool in tabular sheets, they commonly form polygonal, columnar joints (Fig. 5*c*); but if they are extruded or flow into water, they may form pillow structures—tubular masses that are elliptical in cross section (Fig. 5*d*).

Other relatively small volcanic structures include craters, small lava flows, bombs, xenoliths, autoliths, and flow bands. Craters are the circular holes formed by eruption of magma. They range from a few meters to about 1 km (0.6 mi) across. Bombs are masses of lava of various shapes that have been blown into the air and have fallen to the ground as volcanic rock. Xenoliths (foreign stones) and autoliths (earlier crystallized parts of a magma) occur as inclusions in volcanic rocks. Flow bands are millimeter to centimeter thick layers that form during flow from concentrations of bubbles, crystals, or magma of a different composition. Vesicles (holes) in lava flows and amygdules (holes filled with minerals) are considered structures by some geologists. *See* XENOLITH.

Plutonic structures. Plutonic structures occur as both large- and small-sized features in magmas that intrude into rocks of the crust. The large features include batholiths, stocks, lopoliths, roof pendants, cupolas, and various types of dikes and sills. Batholiths are bodies of plutonic rock that are exposed at the surface over an area of 100 km^2 (40 mi^2) or more. Stocks are plutonic bodies of smaller dimension (**Fig. 6**). Batholiths are commonly lens-shaped. Lopoliths are special types of batholiths that are dish-shaped and typically consist of layers of silica-poor rock. Roof pendants are masses of country rock (the rock surrounding a batholith or stock) that hang down into the top of the plutonic rock body (the pluton). Cupolas are dome-shaped masses of the pluton that stick up into the country rock (Fig. 6). *See* PLUTON.

Various types of dikes and sills represent masses of magma that intruded and solidified within country rocks. Dikes are crosscutting, typically sheetlike structures that range from centimeters to kilometers of scale (Fig. 6). Some dikes are cone-shaped or cylindrical and are called cone sheets and ring dikes. Short fat dikes are called apophyses. Sills are intruded sheets of rock that parallel the layers in the country rocks (Fig. 6). When sills are relatively short, thick, and convex on their tops, they are called laccoliths.

Small-scale plutonic structures include dikes, xenoliths, and autoliths, as well as layers and foliation. Gabbroic magmas may undergo a process of fractional crystallization in which crystals form and settle to the bottom of the magma body to form layers of alternating composition. Where early formed crystals are dragged along and aligned by flow of the magma, the aligned crystals define foliation, a structure that also occurs in metamorphic rocks. *See* METAMORPHIC ROCKS.

Magmas. In order to form igneous rocks, older rock must melt to form a magma. Usually only part of a rock melts, a process called anatexis. The magma never has the same chemical composition as the original rock, because some minerals melt before others. Generally, the newly formed magma is richer than the rock from which it came in elements such as silicon, aluminum, sodium, and potassium.

Origins. Magmas form in two general regions: at the base of the crust and in the mantle below the crust. Magmas form when rocks melt as temperature rises (thermal melting), as pressure falls (decompression melting), and as water is added to hot rock (flux melting). Flux melting typically occurs where wet ocean crust is carried down into the mantle (subducted) as a result of plate movements, and water from suducting rocks migrates into hot mantle rocks above the subducted crust. The water causes flux melting that creates basaltic magma. *See* EARTH INTERIOR.

Mantle rocks are hot and plastic. In some places they are cooler than in others, and the hotter rocks tend to rise or flow plastically upward. As they rise, the pressure from the weight of the overlying rock becomes lower. The drop in pressure causes decompression melting of mantle rocks, forming basaltic magmas. Decompression melting is most common in the mantle beneath mid-ocean ridges (spreading centers). *See* MID-OCEANIC RIDGE.

Thermal melting occurs primarily at the bottom of the Earth's crust. Hot basaltic magmas rising from the mantle have extra heat. If they form a pool or pond at the base of the crust, they may have enough heat to melt the crust to form a rhyolitic or granitic magma.

Movement and modification. Once magmas form, they tend to rise because they are hotter and lighter than the rocks around them. Basaltic magmas are more fluid than rhyolitic magmas and can move more rapidly. Through fractures, they can move up 100 km (62 mi) from the mantle in one to a few days. Sometimes they move as diapirs—blobs shaped like an upside-down teardrop. Diapirs move more slowly than magmas in fractures because they must force their way through the overlying rocks. *See* DIAPIR.

Rhyolitic/granitic magmas also move through cracks and as diapirs, but because they are thicker (stickier) they move more slowly than basaltic magmas. In many cases, they do not reach the surface. Although they solidify if they move too slowly, calculations show that if rhyolitic/granitic magmas did not solidify, it might take them more than a million years to move 100 km.

As magmas move upward, they are commonly modified in composition. The three most common processes of modification are fractional crystallization, assimilation, and mixing. Mixing occurs where two magmas of different compositions come together at the same time and place. For example, where basaltic magma melts the bottom of the crust to form a rhyolitic magma, these two magma types may mix to form an andesitic magma.

Assimilation occurs if a magma melts or dissolves the rocks on the sides or walls of a diapir, fracture, or magma chamber, or if pieces of the walls fall into the magma and are absorbed. Generally, the chemistry of wall rocks is different from that of the magma, so assimilated material changes the chemistry of the magma. The chemistry of a magma can also change through fractional crystallization, one of several processes of differentiation. In this process, part (a fraction) of the magma crystallizes to form certain minerals. The minerals have a different chemistry than the whole magma, so when they form they are removing particular chemicals from the magma, changing its composition.

When modified magmas crystallize into rocks, they form rocks of compositions different from those that would have formed from the original magmas. Thus, from the two basic magma types—rhyolitic and basaltic—a whole range of rock types may form.

Loren A. Raymond

Bibliography. M. G. Best and E. H. Christiansen, *Igneous Petrology*, 2001; L. A. Raymond, *Petrology: The Study of Igneous, Sedimentary, and Metamorphic Rocks*, 2d ed., 2002; H. Sigurdsson (ed.), *Encyclopedia of Volcanoes*, 2000; J. D. Winter, *An Introduction to Igneous and Metamorphic Petrology*, 2001.

Ignimbrite

A pyroclastic rock deposit formed by one or more ground-hugging flows of hot volcanic fragments and particles, essentially synonymous with pyroclastic-flow deposit, ash-flow tuff, flood tuff, or welded tuff. When coined by the New Zealand geologist P. Marshall in 1935, the term "ignimbrite" referred only to flows with emplacement temperatures high enough to cause the fragments forming the deposit to become adhered together or welded; this restriction no longer applies, and the term now includes any pyroclastic-flow deposit, welded or not.

Ignimbrites are commonly produced during explosive eruptions and are associated with most of the world's volcanic systems. They vary in size by orders of magnitude (10^{-3} to 10^3 km^3 of erupted material) and have chemical compositions that span the entire range commonly exhibited by igneous rocks (basaltic

A stack of ignimbrites erupted from the Valles Caldera forms the extensive Pajarito Plateau in the Jemez Mountains, New Mexico. (*Photograph by R. A. Bailey, USGS*)

to rhyolitic). An ignimbrite can be of any form and size, but most deposits have sheetlike shapes and cover many thousands of square kilometers. The largest ignimbrite deposits tend to have dacitic to rhyolitic compositions and are associated with powerful, caldera-forming explosive eruptions—usually in the distant geologic past—involving huge eruptive volumes. For example, nearly 4800 km^3 appeared 28 million years ago in a single short-duration eruption from La Garita caldera (San Juan Mountains, Colorado) to form the Fish Canyon Tuff. By comparison, the largest eruption in recorded history (1815, Tambora Volcano, Indonesia) produced only about 30 km^3. Large ignimbrite deposits have sufficient volume and thickness to bury the topography of the underlying terrain and construct nearly flat-topped accumulations, forming distinctive volcanic plateaus such as Yellowstone Plateau (Wyoming and Montana), much of North Island, New Zealand, and the region around Valles Caldera (Jemez Mountains, New Mexico) [see **illus.**].

Ignimbrite deposits are characterized by a poorly sorted aggregate of ash (crystals and glass shards) and pumice. In the larger deposits, the pumice fragments may be flattened and stretched to yield ovoid-to-lenticular shapes, reflecting the compaction and welding of the deposit after or during emplacement. The degree of welding depends on several factors, the most important of which are the thickness of the flow, its temperature, and the amount of contained gas; in general, the greatest welding is seen in the lower half of the thickest part of the deposit.

Specialists in ignimbrite recognize a distinction between flow units and cooling units. A flow unit is a layer that represents a single pyroclastic flow, and the boundaries between flow units are marked by differences—sometimes subtle—in grain size, mineral composition, concentration and nature of pumice, and textural and structural features. If several high-temperature flow units are deposited in rapid succession, they may cool as a single cooling unit. *See* CALDERA; IGNEOUS ROCKS; PUMICE; PYROCLASTIC ROCKS; TUFF; VOLCANIC GLASS; VOLCANO.

Robert I. Tilling

Bibliography. R. A. F. Cas and J. V. Wright, *Volcanic Successions: Modern and Ancient*, Chapman & Hall, London, 1988; R. V. Fisher and H.-U. Schmincke, *Pyroclastic Rocks*, Springer-Verlag, Berlin, 1984; P. Francis, *Volcanoes: A Planetary Perspective*, Oxford University Press, Oxford, U.K., 1993; G. Heiken, *An Atlas of Volcanic Ash*, Smithsonian Contributions to the Earth Sciences, Smithsonian Institution, Washington, DC, 1974; G. A. Macdonald, *Volcanoes*, Prentice Hall, Englewood Cliffs, NJ, 1972; W. E. Scott, Volcanic and related hazards, in R. I. Tilling (ed.), *Short Course in Geology*, vol. 1: *Volcanic Hazards*, American Geophysical Union, Washington, DC, 1989; H. Sigurdsson et al. (eds.), *Encyclopedia of Volcanoes*, Academic Press, San Diego, 2000.

Ignition system

The system in an internal combustion engine that initiates the chemical reaction between fuel and air in the cylinder charge by producing a spark. An ignition system for a multicylinder internal combustion engine has three basic functions: (1) to provide a sufficiently energetic spark to initiate the burning of the fuel-air mixture within each cylinder; (2) to control spark timing for optimum efficiency so that cylinder pressure reaches its maximum value shortly after the piston reaches the top of its compression stroke; and (3) to select the correct cylinder fired.

The ignition system consists of the ignition module, ignition coil, distributor, spark plugs, high-voltage wires, and spark-timing control system. The spark timing function is one of the functions

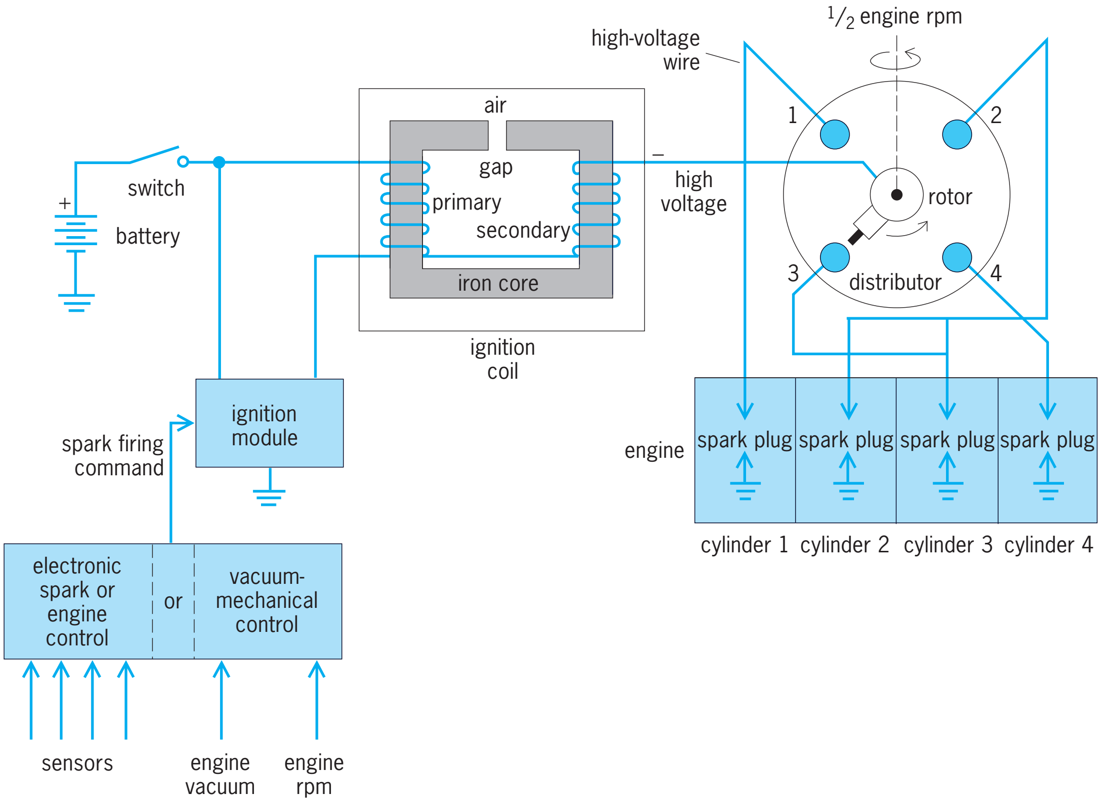


Fig. 1. Inductive discharge ignition system for a four-cylinder engine. The cylinder firing order is 1–3–4–2.

performed by an electronic engine control system or may be performed by a vacuum-mechanical mechanism.

In an inductive ignition system (**Fig. 1**), there are three possible types of control-vacuum-mechanical, electronic spark, or full electronic engine. Prior to spark discharge, electrical energy is stored inductively in the coil primary. The current to the coil primary winding is turned on and off by the ignition module in response to the spark-timing trigger signal. The current-off time marks the beginning of the sparking event. An accurate spark-timing schedule is a complex function of many engine variables, such as fuel-air composition, engine revolutions per minute (rpm), temperature, cylinder pressure, exhaust gas recirculation rate, knock tendency, and engine design. Optimum spark timing is determined from experimental engine testing. A closed-loop electronic feedback control system may dynamically adjust spark timing according to sensor feedback data and internal algorithms.

The ignition coil stores electrical energy during the dwell (current-on) period and acts as a transformer at the end of dwell by converting the low-voltage–high-current energy stored in the primary to high-voltage–low-current energy in the secondary. The distributor selects the fired spark plug by positioning the rotor opposite the terminal connected to one spark plug. The plug selected depends on the cylinder firing order, which in turn depends on the engine design. The distributor is driven at one-half engine speed from the camshaft. *See* SPARK PLUG.

When high voltage (10–30 kV) is created in the coil secondary, a spark jumps from the rotor to a distributor cap terminal, establishing a conducting path from the ignition coil high-voltage terminal along a high-voltage wire to the spark plug. Each cylinder usually has one spark plug. (High-efficiency engines may have two spark plugs per cylinder and two complete ignition systems.) The plug electrodes project as far into the cylinder as possible. After high voltage is applied to the plug, an electrical discharge is generated between its two electrodes. The energy and temperature of this discharge must be sufficient to reliably ignite the fuel-air mixture under all encountered conditions of composition, temperature, and pressure. Otherwise, misfires occur, leading to loss of engine power, excessive emissions, and unsteady power delivery. Reliable ignition results in a self-sustaining flame that travels quickly and completely across the entire cylinder, consuming all the combustible mixture. *See* FLAME.

Inductive discharge ignition. Figure 1 is an example of a standard inductive discharge ignition system. (The term inductive refers to the method of energy storage, which in this case is in the primary coil winding.) Electrical energy is obtained from a 6- or 12-V storage battery, which is connected to the ignition coil primary winding, consisting of several hundred turns of large-diameter wire around an iron core. Wound on the same core are several tens of thousands of turns of small-diameter wire constituting the secondary winding. Energy is stored in the magnetic core with primary current present. When the primary current (1–6 A) is rapidly switched off by the electronic module, current is immediately created in the secondary, since the stored magnetic energy cannot change instantaneously. The ratio of the initial secondary current to the initial primary current is equal to the ratio of primary to secondary turns. The secondary current charges up any capacitance in the secondary circuit, resulting in high spark plug voltage and the creation of a spark. An inductive discharge ignition system has a duration of 0.001 to 0.002 s and a spark energy of 0.005 to 0.050 joule.

In simple ignition systems, the control of spark timing is based only on engine rpm and manifold vacuum. Timing is advanced to occur earlier in the compression stroke as the value for rpm increases and as manifold vacuum increases. Higher manifold vacuum results in lower cylinder pressure during compression and slower combustion. The rpm advance compensates for the ignition delay period occurring between spark firing and the start of rapid combustion. The spark-timing control in vacuum-mechanical ignition systems is accomplished by a combined centrifugal and vacuum advance mechanism that phase-shifts the ignition trigger signal as required.

Other ignition systems. Among the several other types of ignition systems for internal combustion engines are capacitive discharge, multiple-firing capacitive discharge, continuous sustaining, magneto, and distributorless ignitions. The input energy for capacitive discharge systems is stored on a capacitor at several hundred volts (generated by a dc-dc converter). A semiconductor switch (thyristor) controls the discharge of the capacitor into the primary winding. In a multiple-firing capacitive discharge ignition, the ignition module repetitively fires a capacitive discharge ignition during one spark event, increasing both the energy and effective time duration of the spark. In a continuous sustaining ignition, supplemental electrical power is added to the spark after it is established, resulting in electronically controlled extended duration rather than uncontrolled duration as for conventional ignitions. In a magneto ignition, electric current and energy are generated in the primary by relative rotational motion between a magnet and a coil (electromagnetic induction). High voltage is generated in the secondary when a set of contacts in the primary circuit is mechanically opened. Magnetos require no external source of electrical power.

The distributorless ignition system eliminates the need for mechanical distribution of spark energy by using a single coil for one, two, or four cylinders. For the two-cylinder–single-coil system (**Fig. 2***a*),

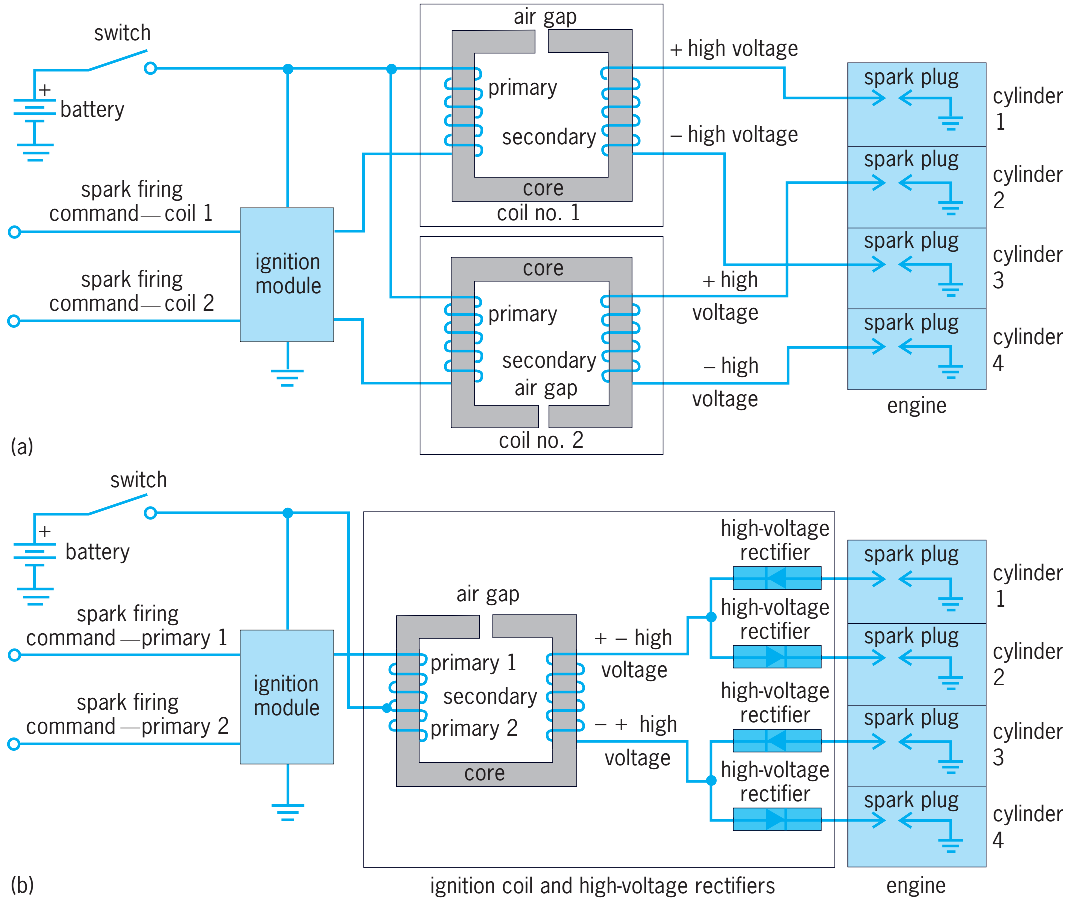


Fig. 2. Distributorless inductive discharge ignition systems for four-cylinder engines. (*a*) Design with two double-ended coils. (*b*) Design with one double-ended coil and four high-voltage rectifiers.

a double-ended ignition coil simultaneously fires a cylinder in a compression stroke together with a second in an exhaust stroke. The exhaust stroke cylinder accepts the waste spark to complete the electrical circuit through the engine block. A design variation (Fig. 2*b*) uses alternating polarity high voltage from a special type of double-ended coil and four high-voltage rectifiers to fire four plugs. The rectifiers steer the voltage to the correct pair of plugs.

Preignition and knock. These are two ignition-related problems that may adversely affect the efficiency of spark ignition engines and even cause damage. Preignition occurs when a hot spot in the cylinder, rather than the spark, initiates combustion, resulting in uncontrolled firing. The hot spot is associated with either cylinder deposits or the spark plug tip itself. Knock is a very rapid, explosive type of combustion that may result in engine damage if severe enough. Knock is often signaled by an audible thudding or pinging. It occurs when an unburned fraction of the premixed fuel-air burns abnormally rapidly. For good efficiency, engines are often operated at light knock (pinging) during low-power operation without incurring damage. *See* SPARK KNOCK.

Compression ignition. In diesel or compression ignition engines, sparkless ignition occurs almost immediately after fuel injection into the cylinder due to high in-cylinder air temperatures. High temperatures result from the high compression ratio of diesel engines. Mechanical or electronic injection timing systems determine ignition timing. *See* COMBUSTION CHAMBER; DIESEL ENGINE; INTERNAL COMBUSTION ENGINE.

Joseph R. Asik

Bibliography. E. A. Avallone and T. Baumeister III (eds.), *Marks' Standard Handbook for Mechanical Engineers*, 10th ed., 1996; D. G. Fink and H. W. Beaty (eds.), *Standard Handbook for Electrical Engineers*, 14th ed., 1999; W. B. Ribbens and N. P. Mansour, *Understanding Automotive Electronics*, 5th ed., 1998.

Illiciales

An order of flowering plants, division Magnoliophyta (Angiospermae), in the subclass Magnoliidae of the class Magnoliopsida (dicotyledons). The order consists of two families, Illiciaceae and Schisandraceae, with fewer than a hundred species in all. They are all woody, with scattered spherical cells containing volatile oils. The leaves are alternate and simple. The flowers are solitary or a few clustered together, regular, and hypogynous; the perianth has five to many segments that are not clearly differentiated into sepals and petals. The Illiciales are related to the Magnoliales and have sometimes been included in that order, but they are apparently more advanced in having fundamentally triaperturate pollen. *Illicium*, with about 40 species in southeastern Asia and in the Caribbean region, is the source of the spice star anise. *See* MAGNOLIIDAE; MAGNOLIOPSIDA.

T. M. Barkley

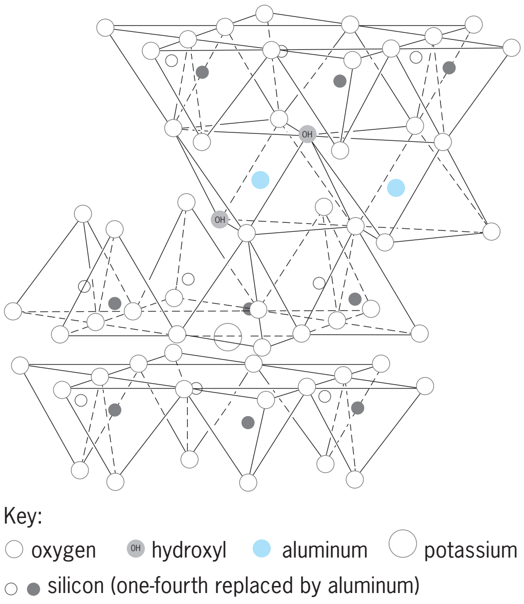
Illite

A clay-size, micaceous mineral; a common component of soil, sediments, sedimentary rocks, and hydrothermal deposits. Generally, the term refers to any relatively nonexpanding clay that possesses a *c* dimension of approximately 1.0 nanometer. Specifically, illite is considered to possess a smaller layer charge and potassium (K) content than muscovite, its potassium (K) content generally being about 0.8 equivalent per half-unit cell as opposed to a content of −1.0 equivalent for muscovite. It is characterized by the ideal formula $Al_2(Si_{3.2}Al_{0.8})O_{10}(OH)_2K_{0.8}$.

Structure. Illite possesses a micalike structure (see **illus.**) consisting of an octahedral sheet sandwiched between two tetrahedral sheets. Apical oxygens from the tetrahedra replace hydroxyls in the alumina octahedra. Negative charge on the 2:1 layer may arise from isomorphous substitutions in both the tetrahedral and octahedral sheets, but tetrahedral substitutions generally predominate. This layer charge is electrically balanced by interlayer potassium ions which, as in muscovite, are fixed in hexagonal cavities between adjacent 2:1 layers in approximate 12-fold coordination with oxygen. Illites may be either dioctahedral or trioctahedral, but the dioctahedral variety is most common. It has been found in five polytypes: disordered one-layer monoclinic ($1M_d$), one-layer monoclinic (1M), two-layer monoclinic ($2M_1$ and $2M_2$), and three-layer triclinic (3T).

Illite frequently is interlayered with smectite, and can be considered the nonswelling end member in an illite-smectite mixed-layer clay series. Pure end-member illite, with no interlayer smectite, is rare.

Properties. Pure illite is characterized by a cation exchange capacity of about 10 milliequivalents/100 g, this capacity arising from broken bonds on crystal edges. Because most illites contain some



Structure of illite. (*After R. E. Grim, Clay Mineralogy, McGraw-Hill, 2d ed., 1968*)

smectite layers, however, measured cation exchange capacities are generally greater than this figure.

Fixed interlayer potassium can be leached with difficulty from illite and replaced by cations of greater hydration energy, thereby transforming illite into vermiculite. This structure can be converted back into illite by adding potassium. This reconstitution can occur during crop fertilization or when clays are transported into the ocean. *See* VERMICULITE.

Formation. Illite forms abundantly by the reaction of smectite during burial diagenesis in sedimentary basins such as the Gulf of Mexico. The reaction is detectable when burial temperature reaches 140°F (60°C); at this depth smectite begins to react to form mixed-layer illite-smectite by dissolving potassium feldspar and fixing interlayer potassium. The reaction is K-feldspar + smectite → illite-smectite + chlorite + quartz. The extent of this reaction is a function of temperature, pressure, solution composition, and time. Illites also can form in potassium-rich weathering environments and by hydrothermal action. An illite like structure also can be formed by subjecting smectite to wetting and drying cycles in the presence of potassium compounds, and by saturating high-charge smectites or vermiculites with K^+. *See* DIAGENESIS; WEATHERING PROCESSES.

Uses. Illitic clays are used for manufacturing structural clay products such as brick and tile. Some degraded illites (vermiculites) are used for molding sands. Illite may also be useful for storing certain types of radioactive wastes, because it is less subject to transformation by heat than are other common clays and because it is highly specific for the sorption of cesium. *See* CLAY, COMMERCIAL; CLAY MINERALS.

Dennis D. Eberl

Bibliography. S. W. Bailey (ed.), *Reviews in Mineralogy*, vol. 13, 1984; G. W. Brindley and G. Brown (eds.), *Crystal Structures of Clay Minerals and Their X-Ray Identification*, 1980; R. E. Grim, *Clay Mineralogy*, 2d ed., 1968.

Illuminance

A term expressing the density of luminous flux incident on a surface. This word has been proposed by the Colorimetry Committee of the Optical Society of America to replace the term illumination. The definitions are the same. The symbol of illumination is E, and the equation is $E = dF/dA$, where A is the area of the illuminated surface and F is the luminous flux. *See* ILLUMINATION; LUMINOUS FLUX; PHOTOMETRY.

Russell C. Putnam

Illumination

In a general sense, the science of the application of lighting. Radiation in the range of wavelengths of 0.38–0.76 micrometer produces the visual effect commonly called light by the response of the average human eye for normal (photopic) brilliance levels. Illumination engineering pertains to the sources of lighting and the design of lighting systems which distribute light to produce a comfortable and effective environment for seeing. In a specific quantitative sense, illumination is the combination of the spatial density of radiant power received at a surface and the effectiveness of that radiation in producing a visual effect.

Illuminance, the alternate term for illumination, is receiving growing worldwide acceptance, particularly in science, but it is not generally used by illuminating engineers in the United States. *See* ILLUMINANCE.

Subjective color. Subjective evaluation of the surfaces of objects as viewed by the human eye has been treated in three classifications. All three are usually grouped by the nonspecialist as “color.” *See* COLOR VISION.

One of the classifications is the attribute of brilliance or subjective brightness, in respect to which every subjective evaluation of the seeing process may be classified according to its subjective intensity, ranging from very dim to very bright.

A second classification is hue, which describes color according to the common color names, such as red, yellow, and green.

The third classification is called variously saturation or purity or vividness of hue. Two colors may be subjectively evaluated as having the same brilliance and the same hue, but the first may seem to contain more white than the second. Conversely, one would evaluate the second color as being a more vivid or pure color than the first.

The hue designation of pure (completely saturated) spectral colors may be approximately assigned to regions of the wavelength spectrum.

Spectral response function of human eye. The spectral response of the human eye to radiation at normal (photopic) brilliance levels was tested by K. S. Gibson and E. P. T. Tyndall. The data from these tests were the source from which the Commission Internationale de l’Eclairage in 1924 established by international agreement a standard brilliance response of the human eye to radiant flux. The function was first known as the CIE values of visibility. The letters ICI (International Commission on Illumination) for the English name are also used. Later the function was named the relative luminosity function, but is now called the spectral luminous efficiency of radiant flux or simply the spectral luminous efficiency. The ordinates of this function are plotted against the wavelength of radiation in **Fig. 1** and are shown in the **table**. *See* LUMINOUS EFFICIENCY.

At threshold (scotopic) levels of seeing, the spectral response function of the human eye is shifted toward lower wavelengths. This is known as the Purkinje effect.

Photometric definitions. Illumination (or illuminance) is the density of radiant power incident upon a receiving surface evaluated throughout the spectrum in terms of the spectral luminous efficiency. Illumination is symbolized by E and is the density of luminous flux per unit of surface area. The unit of illumination in the International System of Units

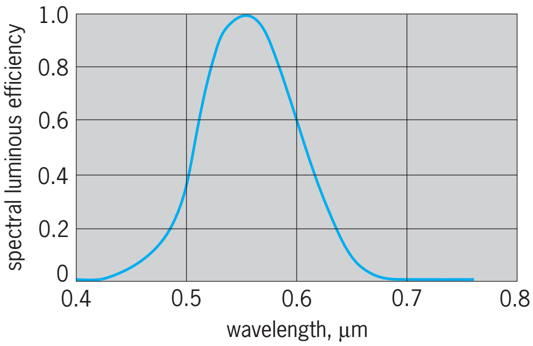


Fig. 1. Spectral response of human eye to radiation. (*After W. B. Boast, Illumination Engineering, 2d ed., McGraw-Hill, 1953*)

(SI) is the lux, which is equal to the lumen per square meter (1 lx = 1 lm/m^2). Units of illumination for other area measures are the phot (ph), equal to lumen per square centimeter, and the footcandle (fc), equal to lumen per square foot. *See* UNITS OF MEASUREMENT.

For a line-spectrum component of irradiation, a simple multiplication of the irradiation component G_i and the value of the spectral luminous efficiency υ_i taken at the proper value of wavelength i yields the component illumination. For continuous-spectrum irradiation, G_λ is a function of wavelength; therefore, the product of the two functions υ_λ and G_λ must be integrated. Equation (1) states mathematically a combination of line spectra and a continuous-spectrum source. Here G_i is the irradiation due to the ith component of the line spectra, υ_i is the spectral luminous efficiency at the wavelength of the ith component irradiation, n is the number of the component line-spectra radiations, G_λ is the spectral irradiation of the continuous spectrum at λ wavelength, υ_λ is the spectral luminous efficiency at the corresponding wavelengths, and k is a constant equal to 683 lm/W.

$$E = k\sum_{i}^{n} \upsilon_i G_i + k\int_0^\infty \upsilon_\lambda G_\lambda \, d\lambda \qquad (1)$$

Spectral luminous efficiency of radiant flux (value of unity at 0.554 μm wavelength)

Wavelength, μm	Efficiency
0.38	0.00004
0.39	0.00012
0.40	0.0004
0.41	0.0012
0.42	0.0040
0.43	0.0116
0.44	0.023
0.45	0.038
0.46	0.060
0.47	0.091
0.48	0.139
0.49	0.208
0.50	0.323
0.51	0.503
0.52	0.710
0.53	0.862
0.54	0.954
0.55	0.995
0.56	0.995
0.57	0.952
0.58	0.870
0.59	0.757
0.60	0.631
0.61	0.503
0.62	0.381
0.63	0.265
0.64	0.175
0.65	0.107
0.66	0.061
0.67	0.032
0.68	0.017
0.69	0.0082
0.70	0.0041
0.71	0.0021
0.72	0.00105
0.73	0.00052
0.74	0.00025
0.75	0.00012
0.76	0.00006

Colorimetry. The experiments from which the curve of Fig. 1 was obtained involved matching the brilliance of radiant flux from narrow spectral bands of adjacent wavelengths by 52 observers. Many scientists had concluded that by adding the radiations from three distinctly different spectral sources properly distributed in the visible spectrum, combinations could be obtained that would uniquely define measurements for all colors. The science of such measurements is called colorimetry. *See* COLOR.

Luminous flux. As previously implied, the lumen is the SI unit of luminous flux that results when radiant power is evaluated through the spectral luminous efficiency function. Related to illumination (or illuminance), the luminous flux Φ received over an area A is given by Eq. (2). The lumen is the only presently used unit of luminous flux and is the SI standard. In Eq. (2) the units E and A must be consistent combinations between lux and square meter, phot and square centimeter, or footcandle and square foot. If E is uniform over the area A, then $\Phi = EA$. *See* LUMINOUS FLUX.

$$\Phi = \int_A E \, dA \qquad (2)$$

Luminous intensity. The luminous intensity I of a source is the radiant intensity J of that source evaluated throughout the spectrum in terms of the spectral luminous efficiency (υ_i or υ_λ or both) in the same manner as illumination was derived from radiant power density upon a receiving surface in Eq. (1).

The concepts of radiant intensity and luminous intensity are strictly applicable only to a point source. In practice, however, flux (radiant or luminous) emanating from a source whose dimensions are negligibly small in comparison with distances from which it is observed may be considered as coming from a point.

The luminous intensity I of a small source, when viewed in a particular direction, is related to the differential of luminous flux $d\Phi$ emitted in that direction in a differential solid angle $d\omega$ by Eq. (3). The

$$I = \frac{d\Phi}{d\omega} \qquad (3)$$

candela, the SI unit of luminous intensity, is related to the lumen as 1 candela equals 1 lumen per unit solid angle (often called steradian).

Quantitative standardization in the field of illumination has been established in various manners by international agreement over the years. At a Paris meeting in October 1979, the General Conference on Weights and Measures redefined the base SI unit candela as the luminous intensity, in a given direction, of a source that emits monochromatic radiation of frequency 540×10^{12} Hz and of which the radiant intensity in that direction is 1/683 watt per steradian (W/sr).

The frequency of 540×10^{12} Hz corresponds to a wavelength of 0.555 μm in free space where the velocity of light is essentially 3×10^8 m/s (1.86×10^5 mi/s). This frequency and wavelength is essentially that where the peak of unity occurs for the spectral luminous efficiency function of Fig. 1. Thus Eq. (4) relates J_i in W/sr for the monochromatic radiation to the resulting monochromatic intensity I_i in candela. The constant k is 683 lm/W as before in Eq. (1). For the source defining the candela, v_i is unity at $f = 540 \times 10^{12}$ Hz, and J_i is 1/683 W/sr, giving Eq. (5). Consistency thus exists among Eqs. (3) and (4) and the constant k from Eq. (1). *See* LUMINOUS INTENSITY; PHYSICAL MEASUREMENT.

$$I_i = kv_iJ_i \tag{4}$$

$$\begin{aligned} I_i = kv_iJ_i &= \left(683\frac{\text{lm}}{\text{W}}\right)(1.000)\left(\frac{1}{683}\frac{\text{W}}{\text{sr}}\right) \\ &= 1\frac{\text{lm}}{\text{sr}} = 1 \text{ candela} \end{aligned} \tag{5}$$

Luminance. Many sources of luminous flux are not small relative to the distance between the luminous source and points on receiving surfaces where illumination results. Consequently, another concept called luminance, or photometric brightness, is very convenient.

The luminance at a point on a radiating surface, acting either from self-emission or by reflection of luminous flux, can be related to a differential segment of the radiating surface dA and the differential luminous intensity dI when the differential area is observed from a particular direction by Eq. (6). The angle α is that between the particular direction of observation and the normal to the differential surface area dA. The denominator $\cos\alpha\ dA$ is the projected area of the differential element from the viewed direction. The equation states that the luminance L is the ratio of the differential luminous intensity to the projected differential area from the direction of observation.

$$L = \frac{dI}{\cos\alpha\ dA} \tag{6}$$

The metric unit of candela per square meter (cd/m^2) has been named the nit (nt) and is recognized as the SI unit of luminance. The fractional value (also metric) of cd/cm^2 is called the stilb (sb). Additional units involving a factor π are the apostilb (asb = cd/π m^2, lambert (L = cd/π m^2), and the footlambert (ftL = cd/π ft^2). The reason for the π factor in these units is discussed below. *See* LUMINANCE.

Inverse-square law. Consider a differential area dA of a luminous surface possessing a differential intensity dI in a specific direction. At a distance D in that specific direction, the differential illumination dE_P upon a receiving surface at a point P is given by Eq. (7), where β is the angle between the normal to the receiving surface and D. Equation (7) indicates that the differential illumination varies inversely as the square of the distance between source and receiving surface. Substituting dI from Eq. (6) into Eq. (7) yields Eq. (8). The total illumination from an extended surface may be obtained by integrating the right-hand side of Eq. (8) over the surface.

$$dE_P = \frac{dI}{D^2}\cos\beta \tag{7}$$

$$dE_P = \frac{L\ \cos\alpha\cos\beta\ dA}{D^2} \tag{8}$$

If the total luminous source is negligibly small relative to D, the inverse square law is given by Eq. (9), where I is the intensity in the direction of D toward P. *See* INVERSE-SQUARE LAW.

$$E_P = \frac{I}{D^2}\cos\beta \tag{9}$$

Light sources. Nature's source of radiation, the Sun, produces radiant power on the Earth extending from wavelengths below 0.3 μm in the ultraviolet region to well over 3 μm in the infrared region of the spectrum. The Sun's spectral radiation per unit of wavelength is greatest in the region of 0.4–0.9 μm. The response of the human eye is well matched to this range of wavelengths. *See* SOLAR RADIATION; VISION.

Oil-flame and gas lighting were used before the advent of the electric light, but since about 1900 electric energy has been the source of essentially all modern lighting devices.

Incandescent lamps. These electric lamps operate by virtue of the incandescence of a filament heated by electric current. The filament usually is composed of tungsten and in the conventional lamp is contained in either a vacuum or inert-gas-filled glass bulb. More recent designs use a halogen regenerative cycle. Currently, iodine or bromide vapors are used in a quartz bulb, which in some designs is internal to an outer, conventional inert-gas-filled bulb. The evaporated tungsten reacts chemically to form a halogen compound, which returns the tungsten to the filament. Bulb blackening is prevented, higher-temperature filaments with increased luminous efficacy result, and higher luminous output can be achieved in relatively small bulbs. Much more efficient optical designs can be achieved for projection equipment, where the source size of light should be small for efficient utilization of the luminous flux. *See* INCANDESCENT LAMP; INFRARED LAMP.

Vapor lamps. These operate by the passage of an electric current through a gas or vapor. In some lamps the

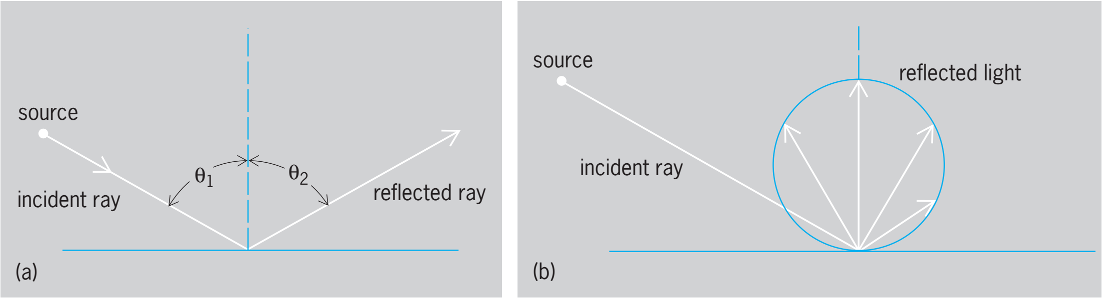


Fig. 2. **Ideal reflection. (*a*) Specular, such as from a polished surface. (*b*) Diffuse, such as from a that surface of microscopic roughness. (*After W. B. Boast, Illumination Engineering, 2d ed., McGraw-Hill, 1953*)**

light may be produced by incandescence of one or both electrodes. In others the radiation results from luminescent phenomena in the electrically excited gas itself. *See* ARC LAMP; VAPOR LAMP.

Fluorescent lamps. These are usually in the form of a glass tube, either straight or curved, coated internally with one or more fluorescent powders called phosphors. Electrodes are located at each end of the tube. The lamp is filled to a low pressure with an inert gas, and a small amount of mercury is added. The electric current passing through the gas and vapor generates ultraviolet radiation, which in turn excites the phosphors to emit light. If the emission of light continues only during the excitation, the process is called fluorescence. If the materials continue to emit light after the source of excitation energy is removed, the process is called phosphorescence. Phosphors for fluorescent lamps are chosen to accentuate the fluorescent action. *See* FLUORESCENT LAMP.

Reflection and transmission. The control of light is of primary importance in illumination engineering because light sources rarely have inherent characteristics of distribution, brightness, or color desirable for direct application. Modification of light may be provided in a number of ways, all of which may be grouped under the general topics of reflection and transmission. Reflection from a surface and transmission through it each may be classified according to their spatial and spectral characteristics.

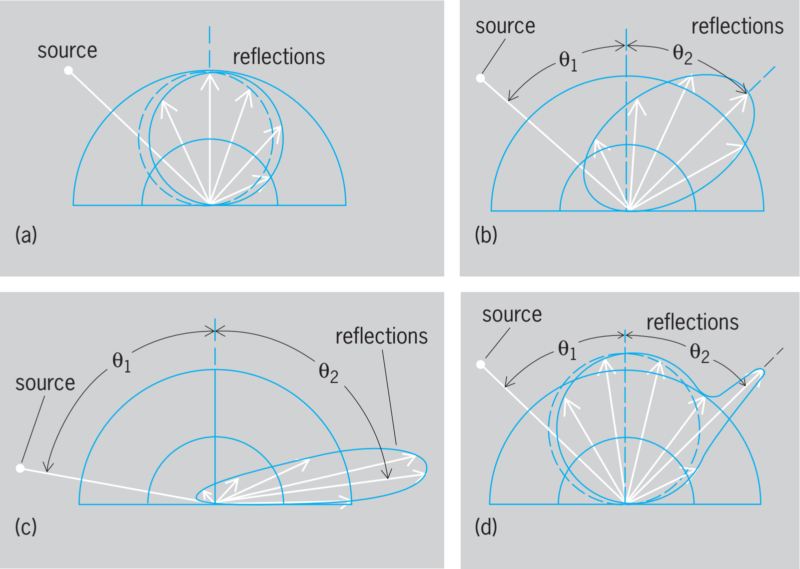


Fig. 3. **Typical reflection characteristics. (*a*) Practical mat surface. (*b*) Semimat surface. (*c*) Semimat surface at large angle of incidence. (*d*) Porcelain-enameled steel. Broken-line curves in *a* and *d* illustrate ideal cosine distribution. (*After W. B. Boast, Illumination Engineering, 2d ed., McGraw-Hill, 1953*)**

Spatial characteristics. Spatially a surface may exhibit reflection conditions ranging from a regular, or specular, reflection to an ideally diffuse characteristic. Similarly, transmission may range spatially from complete transparency to an idealized diffuse transmission.

Figure 2 illustrates the extremes of specular reflection such as would be obtained from polished metal or silvered glass, and of an ideal mat-finished surface possessing microscopic roughness of minute crystals or pigment particles. The specular reflector gives a direct image of the source, with the angle of reflection equal to the angle of incidence (Fig. 2*a*). The plot shown for the diffuse reflector (Fig. 2*b*) illustrates the differential luminous intensity dI of a differential area dA as observed from different directions in the plane of the incident ray. For perfect diffusion, where L is constant for all angles of observation, it follows from Eq. (6) that dI must possess a spatial distribution with α which contains the cos α function. The ideal diffuser illustrated in Fig. 2*b* has such a distribution. The dI distribution with α for a perfect diffuser is the same cosine distribution for all planes passing through the normal to the reflecting surface element dA.

The differential flux density, $d\Phi_E/dA$, emitted by any reflecting surface is symbolized by M and called luminous exitance (luminous emittance is used also but is deprecated). For the perfectly diffusing surface of constant luminance L (called a Lambert surface), differential luminous intensity can be integrated over α from 0 to $\pi/2$ and shown to produce a differential flux $d\Phi_E$ of $\pi L/dA$. The ratio of $d\Phi_E/dA$ is the luminous exitance M. Thus for the perfectly diffusing surface, Eq. (10) is valid.

$$M = \frac{d\Phi_E}{dA} = \pi \qquad (10)$$

The luminance units apostilb, lambert, and footlambert, which contain the factor π in their denominators, are sometimes called rationalized units because they are convenient for expressing the luminance of a perfectly diffusing surface. It is unfortunate that the use of units of luminance established to match for a special situation of perfect diffusion has persisted.

Practical surfaces possess spatial-reflection characteristics intermediate between the ideals of diffuse and specular reflection, particularly at the diffuse end of the range. Typical distribution curves are shown in **Fig. 3**. In all of these illustrations the light source is small. Figure 3*a* demonstrates that even with the best practical mat surfaces, such as dull-finished metals or those painted with flat paint, the location of the source of light has some slight influence upon the intensity distribution as evidenced by an irregularity of the cosine distribution of intensity of the small area of the surface Figure 3*b* and *c* demonstrates the pronounced influence of the location of the light source upon the intensity distribution for surfaces covered with semimat materials, such as satin-finish paint. Figure 3*d* illustrates a mixed-reflection phenomenon resulting when a diffusing surface is overlaid with a surface possessing sheen.

Spatial characteristics of transmission are illustrated in **Fig. 4**. Figure 4*a* shows regular or transparent transmission, such as occurs with clear glass or plastic. Figure 4 *b* shows the diffuse transmission that would result with an idealized diffusing material. In Fig. 4*a* a refraction of the transparent image occurs, but angles θ_1 and θ_2 are equal.

Practical transmission materials possess spatial transmission characteristics intermediate between these ideals and exhibit intensity distribution curves illustrated by **Fig. 5**. Figure 5*a* demonstrates that a perfect cosine intensity distribution is not obtained even with a solid opal- or milk-glass material. A more direct transmission is shown in Fig. 5*b* for a flashed opal-glass medium, in which the majority of the base material is clear glass. Sandblasted or frosted-glass material is much less diffusing, and the location of the light source greatly influences the intensity distribution curves of Fig. 5*c* and *d*.

Spectral characteristics. If a homogeneous radiation of wavelength λ impinges upon a surface, the reflection characteristic or the transmission characteristic will result geometrically according to the spatial considerations of the preceding section. The magnitude of total power density reflected, or transmitted, to that received is defined as the spectral reflectance ρ_λ or the spectral transmittance τ_λ respectively. Thus $\rho_\lambda = J_\lambda/G_\lambda$ and $\tau_\lambda = J'_\lambda/G_\lambda$, where J_λ is the reflected spectral emission at wavelength λ, J'_λ is the transmitted spectral emission at wavelength λ, and G_λ is the spectral irradiation at wavelength λ.

A device for measuring ρ_λ or τ_λ is called a spectrophotometer. Comparison of J_λ or J'_λ with G_λ (or its essential equivalence) may be accomplished by visual or photoelectric methods of photometry.

Either continuous- or line-spectra sources of irradiation may constitute a practical source of radiant-power density. The resulting secondary source of radiant-power density J_λ or J'_λ will be modified, compared with the original, depending on spectral reflection or transmission of the reflecting or transmitting material.

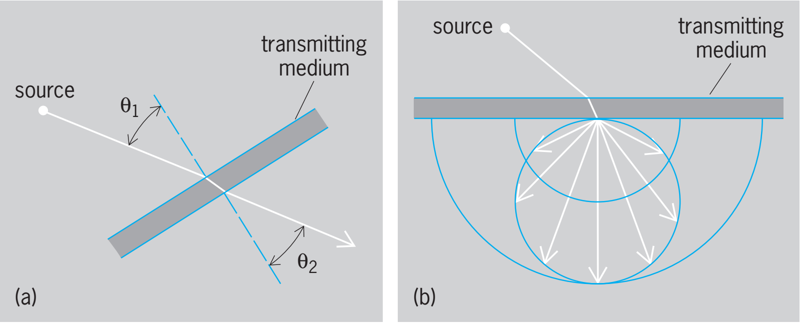


Fig. 4. Spatial characteristics of ideal transmission media. (*a*) Specular. (*b*) Diffuse. (*After W. B. Boast, Illumination Engineering, 2d ed., McGraw-Hill, 1953*)

An evaluation of J_λ or J'_λ using the spectral luminous efficiency ν_λ may be effected for the secondary sources of radiant-power density in the same manner as is done for incident radiant-power density conditions. Both line- and continuous-spectrum effects may be included as demonstrated for the evaluation of the illumination. The evaluation of the reflected or transmitted luminous flux density yields the luminous exitance, symbolized by M for the reflecting surface and M' for the transmitting surface. The functional operations are shown graphically in the series of curves of **Fig. 6** for the reflection resulting from a continuous-spectrum source. The ratio of M or M' to the illuminating E results in $\rho = M/E$ and $\tau = M'/E$, where ρ is the reflectance of the surface and τ is the transmittance of the material. The reflectance and transmittance are dependent upon the source

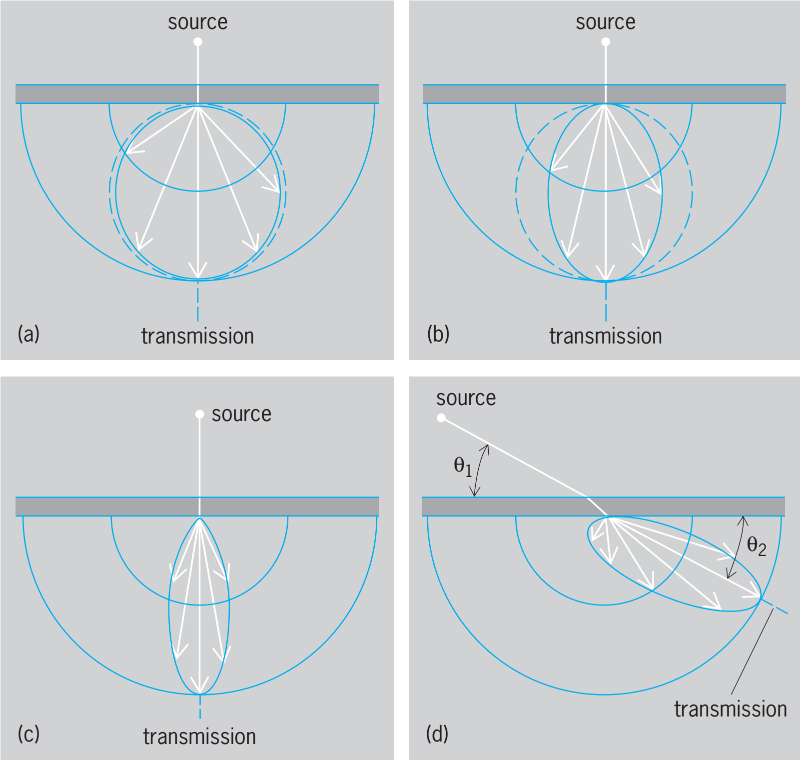


Fig. 5. Transmission characteristics. (*a*) Solid opal glass. (*b*) Flashed opal glass. (*c*) Sand-blasted glass. (*d*) Sand-blasted glass, incident light at 30° from plane at surface. (*After W. B. Boast, Illumination Engineering, 2d ed., McGraw-Hill, 1953*)

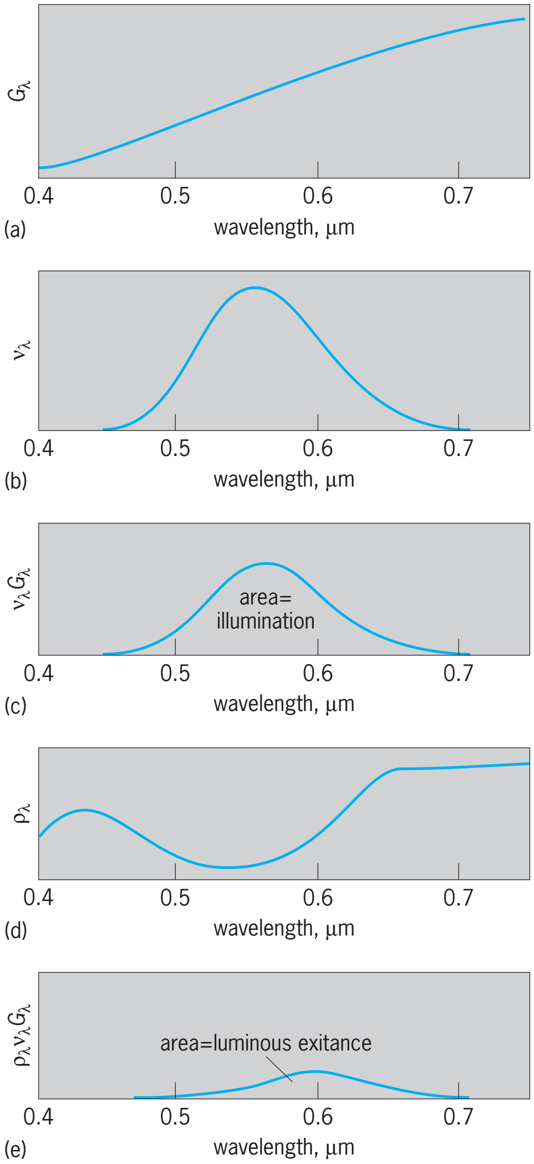


Fig. 6. Process of evaluating the luminous exitance of a reflecting surface after evaluating the illumination. (*a*) Spectral irradiation $G_\lambda$curve. (*b*) Spectral luminous efficiency v_λ curve. (*c*) Curve of $v_\lambda G_\lambda$. Illumination is the area under this curve, the integration of $v_\lambda G_\lambda$. (*d*) Spectral reflectance $_\lambda$ curve. (*e*) Curve of $\rho_\lambda v_\lambda G_\lambda$. Luminous exitance is the area under this curve, the integration of $\rho_\lambda v_\lambda G_\lambda$. (*After W. B. Boast, Illumination Engineering, 2d ed., McGraw-Hill, 1953*)

of illumination as well as the spectral characteristics of the surface or the transmitting meedium. Reflectances and transmittances of surfaces and media are frequently published for the case of a tungsten-filament light source (often without reference to this source) and are useful as guides and indications of magnitude if the sources of illumination do not deviate too greatly from this special distribution.

Light-control methods. Either a control of the character of the reflecting surface or transmitting material or a control of the primary source of luminous flux determines the degree of light control possible. If a light ray is reflected or transmitted diffusely, the shape of the reflecting or transmitting surface is of little importance, although its size may be of great importance. A larger diffuse surface reduces the brightness of the primary source. If a light ray is reflected from a specular surface, more accurate control of the light can be effected.

Parabolic reflectors. These are probably the most useful of all reflector forms. The rays emerge essentially parallel from such a reflector if the light source is placed at its focal point. Searchlights use parabolic mirror reflectors.

Elliptic reflectors. A relatively large amount of light may be made to pass through a small opening before spreading out by using an elliptic reflector. A light source is placed at one of the focal points of a nearly complete elliptical reflector. A small opening is placed at the second focal point where the light is focused. Such a design is used for the pinhole spotlight. For further discussion of reflectors *see* LIGHT PROJECTOR.

Lens control. Refraction of light rays at the boundary between glass and air may be used to control transmitted light in a manner similar to reflector control. Simple thick lenses may be cut away if the cut-away piece is duplicated in the new equivalent surface. Slight irregularities in the field of the light beam occur. These may be of little consequence for searchlights or floodlights, but they would be undesirable for high-quality picture projection. *See* LENS (OPTICS).

Polarization of light. Light rays emitted by common sources may be considered as waves vibrating at all angles in planes at right angles to the direction of the ray. As they pass through some substances or are reflected from specular surfaces, particularly at certain angles, wave components in some directions are absorbed more than those in other directions. Such light is said to be polarized.

Polarized light may be controlled by a transmitting section of a polarizing material. When the absorbing section is oriented to permit passage of the polarized light, minimum control is exercised. If the absorbing section is rotated 90°, the polarized light is essentially completely absorbed. Because sunlight reflected from water or specular surfaces is highly polarized, glare from such sources is effectively controlled by polarized sunglasses or camera filters. Polarized light is used also for detection of strains and defects in glass or plastics. The strained areas appear as color fringes in the material. *See* POLARIZED LIGHT.

Control of primary source. Control of the source of luminous flux can be achieved by controlling the voltage supplied and thus controlling the magnitude of current for some types of light sources or by controlling the time at which the current is permitted to begin on each half cycle of the electric supply for other types of light sources.

For simple incandescent lamps the control of the magnitude of alternating voltage and current can be controlled by variable transformer-type dimmers. Very small silicon controlled rectifiers (SCRs), a modern solid-state development, can accomplish the same control function with equipment which is so small that a unit can be installed in the same space as would be occupied by a conventional, residential "on-off" switch and with very small inherent heat loss caused by the control itself. SCRs are also used in larger capacities for control of large lighting installations. *See* SEMICONDUCTOR RECTIFIER.

The dimming control of fluorescent lamps or other sources employing a gas or vapor discharge is more involved because of the relatively high striking voltage of the vapor discharge. To achieve a wide range of dimming control, the timing on each half-cycle at which current is permitted to begin can be controlled through electronic means and thus produce dimming to low levels. The complexity and cost of such control equipment, however, is greater than the control equipment needed for incandescent lamp dimming.

Interior lighting design. Attainment of a comfortable visual environment is the ultimate goal of interior lighting design. This requires an adequate distribution of illumination and control of brightness of all visible surfaces to avoid excessive contrasts. The adequate distribution of illumination requires that a sufficient number of lighting sources be installed in a manner that produces an adequate level and a proper distribution of illumination. Brightness is controlled by adequate surface reflection or transmission areas associated with the lighting sources and with proper reflecting surfaces within the room itself. Control should ensure that the brightest surfaces of the room are not more than three times the brightness of the work, and the brightness of the work not more than three times that of the darkest surfaces in the room. When such brightness-ratio limitations are satisfied, glare is absent and a satisfactory visual environment is achieved. Proper application of color in the interior enhances overall appearance.

The function of a building greatly influences the manner in which lighting is applied. A particular visual task may require a certain level of illumination, but the lighting design is also influenced by the economics, appearance, and quality of desired results. For example, application techniques are designated as industrial lighting, commercial lighting, and school lighting. Certain application experience and consumer acceptance assume importance in the methods of achieving the level of illumination and control of brightness contrasts. The *IES Lighting Handbook* treats in detail many of these specialized concepts.

Location and spacing of luminaires. Many interior lighting designs utilize general lighting to distribute luminous flux throughout the room on an imaginary working plane, which is usually 30 in. (77 cm) above the floor. To accomplish a reasonably uniform illumination on this plane, certain limitations in spacing between luminaires and in spacing between luminaires and walls must not be exceeded. The height of luminaires above the plane may be an important factor for some interiors. For indirect-type luminaires the distance from the ceiling to the luminaire must provide a proper distribution of light over the ceiling.

Room coefficients. The shape of a room influences the fraction of the emitted luminous flux that will be received on the working plane. Equations and empirical tabulations have been developed to classify rooms according to their shape. Large, broad rooms are more efficient than small or narrow rooms for transferring luminous flux from the luminaire to the working plane, particularly if the reflectance of the side walls is low.

Lumen method of illumination design. This empirical method of design gives the average illumination on the working plane for a particular number of luminaires arranged in a symmetric pattern in a room. The illumination that may be expected to be maintained in the room is given in Eq. (11), where E is

$$E = \frac{N\phi_L K_u M_f}{A} \tag{11}$$

illumination, N is number of luminaires, ϕ_L is initial luminous flux per luminaire in lumens, K_u is coefficient of utilization, M_f is maintenance factor, and A is area of the room.

Tables of the coefficient of utilization are obtained empirically. Usually these tables pertain to a particular luminaire. The coefficient of utilization is a function of the room coefficient and the reflectances of principally the ceiling and side walls. Also called the flux-of-light method, the lumen method was originally designed as a component method whereby the intensity distribution curve of any luminaire could be resolved into three component curves and each calculated according to an equation similar to that shown. However, modern use of the method is principally through coefficients of utilization for each particular luminaire.

The maintenance factor M_f is a number less than one to account for lamp darkening and the collection of dirt within the luminaire, which reduce the illumination emitted from the luminaire.

Zonal-cavity method. The interreflection of luminous flux within a room is accounted for in the empirical data of the lumen method. An exact mathematical calculation of the distribution of luminous flux within rectangular rooms becomes extremely involved. Computer-derived flux-transfer data have resulted in a design method called the zonal-cavity method, which was the method recognized by the Illuminating Engineering Society (*IES*) as of 1969. Calculation procedures are presented in the *IES Lighting Handbook*.

Luminous architectural elements. The original emphasis in the lumen method was upon discrete luminaires arranged symmetrically within the room. The development of the interreflection method of design presumed surfaces of initial luminance on ceiling, walls, and floor as the sources of light within the room. Ingenuity in the application of both methods permits them to be applied both to designs involving luminaires within the room and to designs incorporating many types of luminous architectural elements, such as louvered ceilings, luminous ceilings, and cove systems of lighting.

Natural lighting. Daylight may be an important factor in building design. For comfortable visual conditions any openings for the admission of daylight are planned in position, size, and shape as carefully as any other part of a building. A window can be treated as any other type of light source and included as a part of the illumination and luminance design according to the conventional design procedures. It is usually necessary to provide electric lighting systems when daylight is not available.

Supplementary lighting. Local lighting, such as spotlighting, is frequently used to supplement general lighting where a high level of illumination is desirable on an isolated task.

Exterior lighting design. The nighttime lighting of exterior areas serves many useful purposes. Such lighting extends into the fields of advertising and recreational pursuits as well as utilitarian fields, such as the lighting of industrial yards, parking areas, airports, streets, and highways. The *IES Lighting Handbook* and many textbooks treat in detail many of these specialized applications.

Recommended standards. Illumination levels have been established by the IES to bring all visual tasks to the same level of visibility for normal observers. Some visual tasks are more arduous than others, and higher illuminations provide a compensating factor in bringing such tasks to a satisfactory level of visibility. Standards are established not only for interior lighting designs but also for exterior systems, including recreational lighting. For tables of recommended illumination levels, consult the references in the bibliography.

Warren B. Boast

Bibliography. D. G. Fink and H. W. Beaty (eds.), *Standard Handbook for Electrical Engineers*, 14th ed., 2000; B. Fitt and J. Thornley, *The Control of Light*, 1993; C. Gardner and B. Hannaford, *Lighting Design: An Introduction*, 1993; Illumination Engineering Society, *IES Lighting Handbook, Reference and Application Volume*, 1993; A. Thumann, *Lighting Efficiency Applications*, 2d ed., 1991; L. H. Watson, *Lighting Design Handbook*, 1990.

Image processing

Manipulating data in the form of an image through several possible techniques. An image is usually interpreted as a two-dimensional array of brightness values, and is most familiarly represented by such patterns as those of a photographic print, slide, television screen, or movie screen. An image can be processed optically, or digitally with a computer. This article focuses on digital processing of an image to perform one of five tasks: enhancement, restoration, compression, image fusion, or automated image analysis.

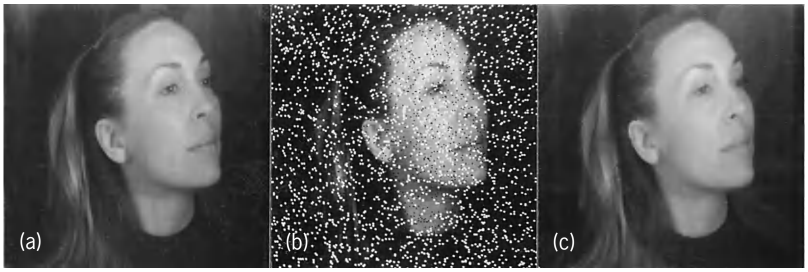


Fig. 1. Noise smoothing. (*a*) Original 256 × 256 pixel image. (*b*) Original image corrupted by spike noise. (*c*) Corrupted image filtered by a 3 × 3 pixel median filter.

Digitization. To digitally process an image, it is first necessary to reduce the image to a series of numbers that can be manipulated by the computer. If, for example, the image is in the form of a photographic transparency, this digitization can be performed by a scanning device that moves a spot of light over the transparency in a rasterlike fashion (from side to side and top to bottom) similar to a television scan. When the light falls on a particular area of the transparency, the amount of light transmitted at that point can be recorded by a photodetector, the signal digitized to, say, eight bits (yielding a possible 2^8, or 256, gray levels), and stored. The spot is then moved a distance usually equal to its width, and the process is repeated, recording the entire transparency in this way as a long string of numbers. The smaller the spot and the finer the scan (up to the limit of the film grain), the more accurate the description of the transparency will be. Each number representing the brightness value of the image at a particular location is called a picture element, or pixel. A typical digitized image may have 512 × 512 or roughly 250,000 pixels, although much larger images are becoming common.

Once the image has been digitized, there are three basic operations that can be performed on it in the computer. For a point operation, a pixel value in the output image depends on a single pixel value in the input image. For local operations, several neighboring pixels in the input image determine the value of an output image pixel. In a global operation, all of the input image pixels contribute to an output image pixel value. These operations, taken singly or in combination, are the means by which the image is enhanced, restored, or compressed.

Enhancement. An image is enhanced when it is modified so that the information it contains is more clearly evident, but enhancement can also include making the image more visually appealing. An example is noise smoothing. Given an image (**Fig. 1***a*), by replacing the value of randomly chosen pixels by 150 (bright gray) or 0 (black), a noisy image (Fig. 1*b*) results. This kind of noise might be produced, for example, by a defective television camera. To smooth the noisy image, median filtering can be applied with a 3 × 3 pixel window. This means that the value of every pixel in the noisy image is recorded, along with the values of its nearest eight neighbors. These nine numbers are then ordered according to size, and the median is selected as the value for the pixel in the new image. As the 3 × 3 window is moved one pixel at a time across the noisy image, the filtered image is formed (Fig. 1*c*). This is a local operation, since the value of a pixel in the new image depends on the values of nine pixels in the old image.

Another example of enhancement is contrast manipulation, where each pixel's value in the new image depends solely on that pixel's value in the old image; in other words, this is a point operation.

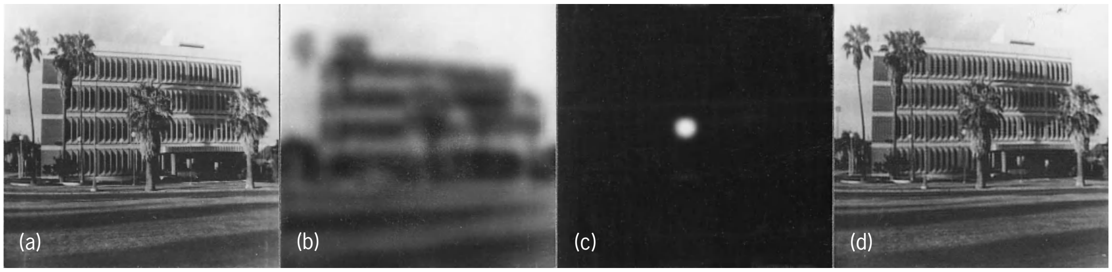


Fig. 2. Example of restoration. (*a*) Original 256 × 256 pixel image (*courtesy of Richard Murphy*). (*b*) Blurred version of the original image, produced by computer simulation. (c) The image of a point subject to the same blurring that produced the blurred version of the original image. (*d*) The results of the inverse filter applied to the blurred version of the original image.

Contrast manipulation is commonly performed by adjusting the brightness and contrast controls on a television set, or by controlling the exposure and development time in printmaking. Another point operation is that of pseudocoloring a black-and-white image, by assigning arbitrary colors to the gray levels. This technique is popular in thermography (the imaging of heat), where hotter objects (with high pixel values) are assigned one color (for example, red), and cool objects (with low pixel values) are assigned another color (for example, blue), with other colors assigned to intermediate values.

Restoration. The aim of restoration is also to improve the image, but unlike enhancement, knowledge of how the image was formed is used in an attempt to retrieve the ideal (uncorrupted) image (**Fig.** 2*a*). Any image-forming system is not perfect, and will introduce artifacts (for example, blurring, aberrations) into the final image that would not be present in an ideal image. A computer can simulate an out-of-focus lens, producing a blurred image (Fig. 2*b*). The image of a point object through this system (Fig. 2*c*) is called the point spread function of the system. Such blurring can be referred to as a local operation, in that the value of a pixel in the blurred image is an average of the values of the pixels around (and including) that pixel location in the original image. The blurring process is mathematically referred to as a convolution of the original object with the point spread function.

Another point spread function, called a filter, can be constructed that undoes the blurring caused by the original point spread function. Such a point spread function has both positive and negative values, even though the blurring point spread function has only positive values. By imaging the blurred image (Fig. 2*b*) with the filter point spread function, the restored image (Fig. 2*d*) results. The filter point spread function is spread out more than the blurring point spread function, bringing more pixels into the averaging process. This is an example of a global operation, since perhaps all of the pixels of the blurred image can contribute to the value of a single pixel in the restored image. This type of deblurring is called inverse filtering, and is sensitive to the presence of noise in the blurred image. By modifying the deblurring filter according to the properties of the noise, performance can be improved. An example of the need to deblur images from an optical system is the Hubble Space Telescope before its spherical aberration was corrected with new optics. *See* SATELLITE ASTRONOMY.

Mathematical transformations can be performed on images to facilitate deblurring, as well as other operations. The Fourier transform decomposes the image into two-dimensional sine waves of varying frequency, direction, and phase. Convolution of an image with a point spread function (the imaging process) can be carried out in the Fourier domain by simply multiplying the Fourier transform of the image by the Fourier transform of the point spread function, and taking the inverse Fourier transform of this product to get the final image. *See* FOURIER SERIES AND TRANSFORMS.

Compression. Compression is a way of representing an image by fewer numbers, at the same time minimizing the degradation of the information contained in the image. Compression is important because of the large quantities of digital imagery that are sent electronically and stored. Digital high-definition television relies heavily on image compression to enable transmission and display of large-format color images.

Compression can be performed in the image space or in transform space (for example, Fourier). In wavelet decomposition, a particularly useful type of transformation, the image is broken (that is, decomposed) into successively smaller representations of itself by a downsizing process, with each representation containing information about the original image at different size scales. The smaller representations are then wavelet-transformed. The wavelet transform reduces the correlation and dynamic range of the image representations, so that after the transformation each representation requires considerably fewer bits to encode the information that it contains. *See* WAVELETS.

All of the smaller representations (also referred to as subbands) are included in the compressed image. The dynamic range of each subband is different; smaller subbands have a larger dynamic range (that is, require more bits per pixel) than larger subbands, and the compression algorithm automatically takes this into account.

Once the image is compressed for storage or transmission, it must be uncompressed for use, by the inverse of the compression operations. There is a trade-off between the amount of compression and


(a) (b) (c)

Fig. 3. Example of compression. (*a*) Original 1024 × 768 pixel image, at 3.3 bits per pixel (bpp), typical of the compression provided by digital cameras. (*b*) Compressed version of the original image using JPEG compression, at 0.7 bpp. There is little visual degradation, yet the file size is over 4 times smaller. (*c*) Further compressing the image to 0.4 bpp; noticeable degradation has occurred with this lossy algorithm.

the quality of the uncompressed image. High compression rates (giving rise to so-called lossy compression) are acceptable with television images, for example. However, where high image quality must be preserved (lossless compression, as in diagnostic medical images), only compression rates as low as three to four may be acceptable. Figure 3 is an example showing the tradeoff between image compression and image quality, for the JPEG compression scheme. *See* DATA COMPRESSION.

Many image formats have been developed for the storage and transmission of image data, and both lossy and lossless compression methods are incorporated into them. Familiar examples of such formats are JPEG (lossy), GIF (lossless LZW algorithm, limited to 256 colors), PNG (lossless, an open-source updated GIF format supporting millions of colors), TIFF (lossy or lossless), and Exif (a common format used with digital cameras that supports JPEG compression).

Image fusion. This term refers to the combination of information from two different images to produce a new image that provides more information to the viewer than either original image by itself. Such combination can provide several advantages: low-resolution infrared information (which provides information about the temperature of an object) can be mapped onto a higher-resolution visible image or a road map to aid in object location and recognition; objects invisible in one image may be visible in the other, so the fused image allows object relationships to be clarified; and the addition of elevation data to a visible, infrared, or radar image allows a three-dimensional stereo view of the scene to be produced. In medicine, low-resolution (5–10-mm) positron emission tomography (PET) data, which represent the functioning of the brain, can be combined with high-resolution (1-mm) magnetic resonance image (MRI) data that represent anatomical information. *See* MEDICAL IMAGING.

A basic requirement of image fusion is that the original images be registered. This usually means that a set of control points must be identified in each set of images that correspond to the same objects. One image is called the reference image, and the other image is geometrically distorted until the position of its control points matches the control points in the reference image.

Simple techniques for image fusion add or multiply the pixel values of the registered images together to form the new image. More sophisticated techniques involve nonlinear combinations of pixel values or involve transformations of the registered images to form the fusion product. The coherent fusion of two synthetic-aperture radar (SAR) images of the ground taken days apart can allow the imaging of the motion of ice flows. *See* RADAR; REMOTE SENSING.

Geometrical morphing of one image to another can be loosely considered another kind of image fusion, in that a temporal progression of one image into another is formed by geometrically distorting one image into another, with corresponding grayscale and color changes to the overlapping pixels. The effect is to make one image flow into another. This is used extensively in the advertisement and entertainment industry, though not as frequently in science and engineering.

Automated image analysis. This technique, also referred to as image understanding and image exploitation, involves automatically extracting information from an image. This information extraction can take many forms, such as identifying a tumor in a medical image; finding a flaw in an integrated-circuit chip; looking for misshapen candy on an assembly line; spotting a small hole in a sheet of metal moving at high speed; identifying an enemy tank, airplane, or missile in milliseconds; or identifying a person's face.

In the automatic extraction of image information, first the image is usually enhanced or restored to facilitate the analysis. Next, measurable quantities, called features, are extracted from the image which represent the structures that are being sought.

One use of the extracted features is to label each pixel in the image, forming a segmentation of the image. A segmentation represents a homogeneous area of the image (that is, an area with similar feature characteristics) as a single color or gray level. Areas with different properties, based upon their features, are represented by different colors. Once all the pixels of an image are labeled in this manner, this new image can be presented to a human observer for a decision, or higher-level automated decision processing can take place. For example, in computer-aided diagnosis in medicine, a tumor in an MRI of the brain can be labeled a different color than the surrounding healthy tissue, based upon subtle textural differences of the tumor tissue. This automatic labeling can assist the radiologist in diagnosis or can be the jumping-off point for further automated processing. *See* CHARACTER RECOGNITION; COMPUTER VISION.

Applications. Image processing is an active area of research in such diverse fields as medicine, astronomy, microscopy, seismology, defense, industrial quality control, security, and the publication and entertainment industries. The concept of an image has expanded to include three-dimensional data sets (volume images), and even four-dimensional volume-time data sets. An example of the latter is a volume image of a beating heart, obtainable with x-ray computed tomography (CT). CT, PET, single-photon emission computed tomography (SPECT), MRI, ultrasound, SAR, confocal microscopy, scanning tunneling microscopy, atomic force microscopy, and other modalities have been developed to provide digitized images directly. Digital images are widely available from the Internet, CD-ROMs, and inexpensive charge-coupled-device (CCD) cameras, scanners, and frame grabbers. Software for manipulating images is also widely available. *See* CHARGE-COUPLED DEVICES; COMPACT DISK; COMPUTER STORAGE TECHNOLOGY; COMPUTERIZED TOMOGRAPHY; CONFOCAL MICROSCOPY; DATA COMMUNICATIONS; DIGITAL COMPUTER; ELECTRON MICROSCOPE; MEDICAL ULTRASONIC TOMOGRAPHY; NUCLEAR MEDICINE; SCANNING TUNNELING MICROSCOPE; TELEVISION CAMERA. Warren E. Smith

Bibliography. G. A. Baxes, *Digital Image Processing: Principles and Applications*, 1994; K. R. Castleman, *Digital Image Processing*, 1996; R. C. Gonzalez, R. E. Woods, and S. L. Eddins, *Digital Image Processing Using MATLAB*, 2003; W. K. Pratt, *Digital Image Processing: PIKS Inside*, 3d ed., 2001; J. C. Russ, *The Image Processing Handbook*, 4th ed., 2002.

Image tube (astronomy)

A photoelectric device for intensifying faint astronomical images. In these devices a photoemissive surface, called the photocathode, emits electrons through the photoelectric effect. In most tubes the photocathode is semitransparent and is deposited on the inside of a transparent window that is mounted on the end of an evacuated glass or ceramic cylinder. When light from a telescope or spectrograph is imaged on the photocathode, electrons are ejected into the vacuum inside the tube. Electric fields or electric and magnetic fields in combination then accelerate and direct the photoelectrons through the device. In an image intensifier tube the photoelectrons are ultimately reimaged on a phosphor-coated output window that converts them back into visible light. An external television camera or charge-coupled device (CCD) is used to record the image. Optical gain is provided by the phosphor, which may release as many as a thousand photons for each photoelectron which crashes into it. In other devices the electron image is projected onto a solid-state diode array or a grid of anodes that is built directly into the tube. *See* CATHODE-RAY TUBE; PHOSPHORESCENCE; PHOTOELECTRIC DEVICES; PHOTOEMISSION; PHOTON.

The quantum efficiency of an image tube is the probability that an incoming photon of light will release a detectable photoelectron inside the tube. Although a single photon of light is all that is necessary to create a photoelectron, some photons pass through the photocathode without being absorbed. Others are absorbed but do not produce a free electron. In spite of these losses, photocathodes with quantum efficiencies above 20% are routinely fabricated. Bialkali and multialkali photocathodes are useful in the optical and ultraviolet and are composed of antimony and two or more alkali metals. Solar-blind photocathodes are made from cesium and iodine or tellurium. They have high quantum efficiencies in the ultraviolet but are insensitive at optical wavelengths. All photocathodes emit a few electrons even in the absence of light. This dark emission is a source of noise and must be taken into account when observing faint sources. Because modern CCDs can have quantum efficiencies over 60% in the optical wavelength range, photoemissive devices have largely been phased out of ground-based astronomy. The extended red sensitivity of CCDs effectively precludes their use in the far-ultraviolet, however, since even the best ultraviolet filters have substantial red leaks. Photoemissive detectors have thus been flown on virtually all ultraviolet astronomy satellites such as the *International Ultraviolet Explorer* and the Hubble Space Telescope. *See* CHARGE-COUPLED DEVICES; ULTRAVIOLET ASTRONOMY.

Electrostatically focused tubes. Each tube in an electrostatically focused image intensifier (**Fig. 1**) is constructed with fiber-optic input and output windows that are flat on the outside and spherical on the inside. Light falling on the front of each detector is piped to the photocathode by the optical fibers of the input window. The ejected photoelectrons then pass through a small hole in the tip of a conical electrode and come to focus on a phosphor screen that has been deposited on the inner surface of the output window. A reflective coating over the phosphor prevents light from traveling back to the photocathode. High gain is achieved by having successive stages of amplification. While electrostatically focused image tubes have low dark emission rates, they have poor distortion characteristics. They are also not useful in the ultraviolet because of the fiber-optic input window. *See* OPTICAL FIBERS.

Proximity-focused tubes. Proximity-focused image intensifiers are very short and have plane-parallel input and output windows. There are no electrodes for imaging the photoelectrons; a strong electric field simply accelerates them into the phosphor before any significant defocusing can take place. While proximity-focused image tubes have almost no image

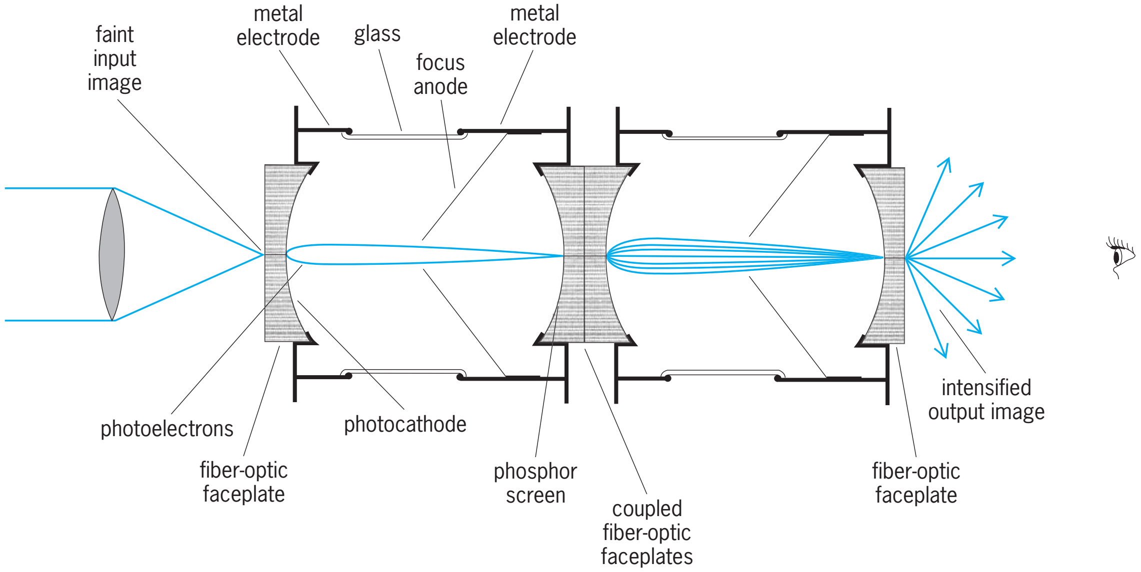


Fig. 1. Two-stage electrostatically focused image tube, fiber-optically coupled.

distortion, they have high dark emission rates and relatively low gain.

Microchannel-plate tubes. A microchannel plate (**Fig. 2**) consists of an array of closely spaced tubes, each of which is coated on the inside with a secondary-emitting material. When a free electron enters a channel, it eventually collides with the tube wall and releases several secondary electrons. These electrons then travel farther down the tube and produce secondary electrons of their own. Since this process is repeated several times down the channel, thousands of electrons exit the rear face of the microchannel plate. A microchannel plate can be added to a proximity-focused image intensifier to obtain higher gain at an acceptable level of dark emission. In such an application, the microchannel plate is simply sandwiched between the photocathode and the phosphor. For space applications, the photocathode is usually deposited directly on the microchannel plate. Since this technique eliminates the photocathode window, such a device can be used in the extreme ultraviolet. In a multianode microchannel plate device, the phosphor screen is replaced by a grid of anodes that records the position of each pulse of electrons that exits the plate.

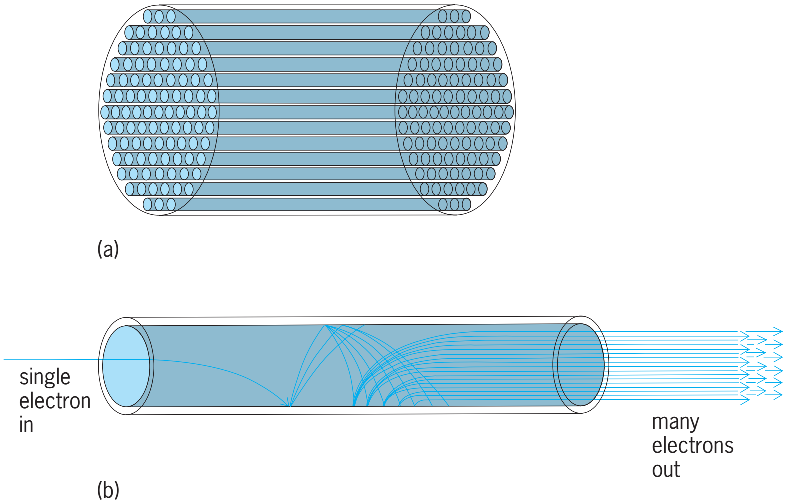


Fig. 2. Microchannel plate showing (*a*) array of tubes and (*b*) details of a single channel.

Magnetically focused tubes. Magnetically focused image tubes use electric and magnetic fields for electron imaging and have plane-parallel input and output windows. Electrons from the photocathode are accelerated down the tube by a series of ring-shaped electrodes. As the electrons are accelerated, an axial magnetic field from an external permanent magnet or solenoid forces the photoelectrons to follow spiral paths that reconverge at the output window of the tube. In an image intensifier the electron image is converted back into light. In other types of tubes the electron image is focused directly on a silicon diode array, and each incoming electron produces a large number of electron-hole pairs in the silicon instead of a burst of photons. While magnetically focused devices have some distortion, they usually have very low dark emission rates. *See* ASTRONOMICAL PHOTOGRAPHY; ELECTRON MOTION IN VACUUM.

Richard G. Allen

Bibliography. A. Boksenberg and D. L. Crawford (eds.), *Proceedings of SPIE*, vol. 445: *Instrumentation in Astronomy V*, 1983; D. L. Crawford (ed.), *Proceedings of SPIE*, vol. 331: *Instrumentation in Astronomy IV*, 1982; P. W. Hawkes (ed.), *Advances in Electronics and Electron Physics*, vols. 64A and 64B, 1985.

Immittance

The impedance or admittance of an alternating-current circuit. It is sometimes convenient to use the term immittance when referring to a complex number which may be either the impedance (ratio of voltage to current) or the admittance (ratio of current to voltage) of an electrical circuit. The units of impedance and admittance are, of course, different and so units cannot be assigned to an immittance. However, in certain theoretical work it may be necessary to deal with general functions which afterward will be specialized to become either an impedance or an admittance by the assignment of suitable units; in such cases it is convenient to refer to the functions as immittances. *See* ADMITTANCE; ALTERNATING-CURRENT CIRCUIT THEORY; ELECTRICAL IMPEDANCE. J. O. Scanlan

Immune complex disease

Local or systemic tissue injury caused by the vascular deposition of products of antigen-antibody interaction, termed immune complexes. Immune complex formation with specific antibodies causes the inactivation of potentially dangerous antigenic substances and their elimination by phagocytic cells. There are harmful consequences only when immune complexes deposit in tissues, inciting various mediators of inflammation, leading to injury and disease. When the reaction takes place in the extravascular fluids near the site of origin of the antigen (by injection, secretion, and such), focal injury can occur, as exemplified by the Arthus reaction or such conditions as experimental immune thyroiditis. Systemic disease may occur when soluble antigens combine with antibodies in the vascular compartment, forming circulating immune complexes that are trapped nonspecifically in the vascular beds of various organs, causing such clinical diseases as serum sickness or systemic lupus erythematosus with vasculitis and glomerulonephritis. The term immune complex disease usually signifies this systemic immune complex formation and vascular deposition. In the widely used classification originated by Gell and Coombs, immune complex disease is Type III immunopathology. *See* ANTIGEN-ANTIBODY REACTION; CONNECTIVE TISSUE DISEASE.

Pathophysiology. Circulating immune complex disease occurs when the host's antibody production, relative to the amount of antigens, is inadequate for prompt elimination of antigen. Normally, excess amounts of antibody are formed which generate large immune complexes that are removed very rapidly from the circulation and are disposed of by the mononuclear phagocytic system. If the antibody response is very poor, only a few very small complexes are formed which are not prone to vascular deposition. When the relative antibody production is such that complexes of intermediate size form, vascular trapping can occur and injury results from the effects of inflammation. In addition to immune complex size, other factors influence vessel deposition, including the efficiency of systemic clearance of immune complexes, the hemodynamics of blood flow, and vasoactive amine-influenced changes in vascular permeability. Through dynamic equilibrium, continual modification of the deposits occurs as antigen and antibody fluctuate in the body fluids.

Human immune complex disease. Circulating immune complexes play a prominent role in several types of human disease. They can follow the production of antibody to exogenous antigens (iatrogenic, infectious, and so on), or to endogenous ones during an autoimmune response. Some individuals with streptococcal infections of the throat or skin develop poststreptococcal glomerulonephritis with evidence of circulating immune complexes, and complex deposition in the kidney. Although usually short-lived, this condition can progress to renal insufficiency. The more chronic infection with hepatitis B virus can result in immune complex disease. A transient serum sickness–like illness may occur as a prodrome to the hepatic illness. In individuals with prolonged circulation of hepatitis B viral antigens, severe immune complex vasculitis and glomerulonephritis may develop. Chronic parasitic diseases such as malaria and schistosomiasis are often complicated by complex-related problems. Foreign serum and drugs used in treatment or illicitly can cause immune complex injury.

Immune complexes are often responsible for some of the systemic manifestations of autoimmune diseases, for example, in systemic lupus erythematosus. Tissue biopsy is used to identify complex deposition. Vasculitis (inflammation of the blood vessel walls) results from the vascular immune complex deposit in individuals with systemic lupus erythematosus as well as in those with other so-called autoimmune diseases such as rheumatoid arthritis and polyarteritis nodosa. In most cases of vasculitis, the offending antigen is unknown, although nuclear antigens including DNA are involved in lupus, as is hepatitis B antigen in some cases of polyarteritis nodosa.

Circulating complexes occur in individuals with autoimmune thyroiditis as well as individuals with neoplasms, most commonly carcinomas. Signs of immune complex tissue damage are infrequent in individuals with neoplasia except for occasional glomerulonephritis in which tumor-associated antigens (for example, carcinoembryonic antigen) have been demonstrated in immune complex deposits in glomeruli.

Because of its unique role as a filter, the kidney is the most common organ in which immune complexes deposit. Most immunologic glomerular injury is considered to be a consequence of complex deposition, and this mechanism is recognized in tubulointerstitial renal disease as well. In some immune complex-mediated glomerular diseases, the relevant antigens are known, as in postinfectious disease, malignancy, and systemic lupus erythematosus.

Diagnosis. The diagnosis of immune complex disease in compatible clinical situations is aided by tests including the detection of complement consumption and the presence of anti-DNA antibodies in systemic lupus erythematosus.

Tissue deposition of complexes is routinely determined by immunofluorescence microscopy using fluorescein conjugated antibodies specific for components of the immune complex, namely various classes of immunoglobulin, complement components, and, when known, the antigen itself. The random deposition of complexes in tissue results in irregular granular collections of immunoglobulin within vessel walls or glomeruli. Although positive immunofluorescent demonstration of granular immunoglobulin and complement deposits in tissue is suggestive, it is not pathognomonic of immune complex localization, unless antigen and antibody can be demonstrated in a similar distribution. Alternately, antibody may be eluted from tissue and tested for antigen specificity. Electron microscopy often reveals electron-dense deposits corresponding to sites of complex deposition, allowing detailed morphologic assessment. Immunoelectron microscopy using labeled antibodies (such as ferritin or peroxidase) may provide additional information as to the precise location of deposited complexes. *See* IMMUNOFLUORESCENCE.

A number of methods have been described which detect circulating immune complexes with varying degrees of success. In general, the assays more often detect complexes in acute forms of disease when the levels of complex are presumably higher than in more indolent disease processes. The tests may be divided into those based on physicochemical properties and on immunochemical properties. Physicochemical methods take advantage of size and configurational changes in complexed immunoglobulin. Ultracentrifugation and column chromatography are relatively insensitive in separating the immune complexes from smaller uncomplexed immunoglobulin, so that large amounts of complexes must be present for detection. Precipitation of immune complexes with poly(ethylene glycol) of appropriate concentration is relatively more sensitive. Cryoprecipitation may also detect some immune complexes.

The immunochemical methods tend to be more sensitive and employ the ability of complexed immunoglobulin to interact with complement, bovine conglutinin, rheumatoid factor, anti-antibody sera or cellular elements. The complement-based assays, most often relying on binding of a portion of the first complement component, Clq, to the immune complexes (in solid-phase or fluid-phase radioimmunoassay) are relatively sensitive. *See* COMPLEMENT; IMMUNOGLOBULIN.

Assays in which immune complexes cause platelet aggregation or bind to immunoglobulin Fc, or complement, receptors on cells have been used quite successfully to detect small quantities of complex in sera. The most common cell line used is the lymphoblastoid Raji cell line derived from a person with Burkitt's lymphoma. This line binds complexes attached by the complement components. The immune complexes are subsequently quantified by radioimmune assay. Advantage may be taken of the different reactivities of the various assays for identification and more importantly isolation and characterization of the immune complex. *See* RADIOIMMUNOASSAY.

Treatment. Treatment of immune complex disease can be divided into nonspecific and more specific modalities. Primary among the specific measures is the identification and elimination of the offending antigen. This may be possible with some infections when specific therapies are available, and in certain instances where the antigenic source can be removed, such as a neoplasm. More frequently, nonspecific anti-inflammatory (corticosteroids) and immunosuppressive agents (such as cyclophosphamide and azathioprine) are used to attempt to blunt the person's immune response, thereby lessening the amount of immune complexes produced. Removal of the circulating plasma by plasmapheresis is also being tried as a way of removing circulatory immune complexes and perhaps circulating mediators of inflammation as well. *See* AUTOIMMUNITY; IMMUNOLOGY.

Edward H. Cole; Curtis B. Wilson

Bibliography. E. Braunwald et al. (eds.), *Harrison's Principles of Internal Medicine*, 15th ed., 2001; B. M. Brenner and F. C. Rector, Jr. (eds.), *The Kidney*, 2 vols., 6th ed., 2000; C. G. Cochrane and D. Koffler, Immune complex disease in experimental animals and man, *Adv. Immunol.*, 16:185–264, 1973; A. N. Theofilopoulos and F. J. Dixon, The biology and detection of immune complexes, *Adv. Immunol.*, 28:89–220, 1979; C. B. Wilson, B. M. Brenner, and J. H. Stein (eds.), *Contemporary Issues in Nephrology*, vol. 3, 1979.

Immunity

A state of resistance to an agent, the pathogen, that normally produces an infection. Pathogens include microorganisms such as bacteria and viruses, as well as larger parasites. The immune response that generates immunity is also responsible in some situations for allergies, delayed hypersensitivity states, autoimmune disease, and transplant rejection. *See* ALLERGY; AUTOIMMUNITY; TRANSPLANTATION BIOLOGY.

Immunity is engendered by the host immune system, reacting in very specific ways to foreign components (such as proteins) of particular parasites or infective agents. It is influenced by many factors, including the environment, inherited genes, and acquired characteristics. Reaction to a pathogen is through a nonadaptive or innate response as well as an adaptive immune response. The innate response is not improved by repeated encounters with the pathogen. An adaptive response is characterized by specificity and memory: if reinfection occurs, the host will mount an enhanced response.

The components of the pathogen that give rise to an immune response, to which antibodies are generated, are called antigens. There are two types of specific responses to an antigen, antibodies and the cellular response. Antibodies help to neutralize the infectious agent by specifically binding it. A series of proteins in the blood (called complement) act in conjunction with antibodies to destroy pathogenic bacteria. In the cellular response, cytotoxic T cells are recruited to kill cells infected with intracellular agents such as viruses. Helper T cells may also be generated, which influence B cells to produce appropriate antibodies. Inflammatory responses and activation of other kinds of cells, such as macrophages, in conjunction with lymphocytes, is another important aspect of the immune response, as in delayed hypersensitivity. This kind of response seems to be common in certain chronic infections. *See* ANTIBODY; ANTIGEN; COMPLEMENT.

Complex immune systems (antibody and specific cellular responses) have been demonstrated in mammals, birds, amphibians, and fish, and are probably restricted to vertebrates. The evolutionary strategy of eliminating diseased individuals is not feasible in vertebrates because they have so few progeny. Many other organisms have self and nonself discrimination mechanisms and primitive defense strategies, most likely lacking the flexibility as well as the specificity and memory of vertebrate systems.

Natural or innate immunity. There are natural barriers to infection, both physical and physiological, which are known collectively as innate immunity, and include the effects of certain cells (macrophages, neutrophils and natural killer cells) and substances such as serum proteins, cytokines, complement, lectins, and lipid-binding proteins. The skin or mucous membranes of the respiratory tract are obvious barriers and may contain bacteriostatic or bactericidal agents (such as lysozyme and spermine) that delay widespread infection until other defenses can be mobilized. Fatty acids provide a good example of natural immunity, such as in foot rot (athlete's foot) in which infection tends to occur between the small toes where there are no sebaceous glands.

If organisms manage to enter tissues, they are often recognized by molecules present in serum and by receptors on cells. Bacterial cell walls, for example, contain substances such as lipopolysaccharides that activate the complement pathway or trigger phagocytic cells. Some receptors on cells, called Toll receptors, are related to proteins found on plants or insects that trigger the production of antimicrobial substances. These innate responses to infectious agents are obviously ancient and may have originated before common ancestors of humans, plants, and insects.

Host range is dramatic in its specificity. Animals and plants are generally not susceptible to each other's pathogens. Within each kingdom, infectious agents are usually adapted to affect a restricted range of species. For example, mice are not known to be susceptible to pneumococcal pneumonia under natural conditions. If a small number of organisms are injected, bypassing the natural barriers, a fatal infection results. Domestic cats never contract measles. The lack of infectivity of rodents with human immunodeficiency virus (HIV) has meant that monkeys have to be used to study the behavior of this retrovirus. In some cases, disparate species lack the receptors at their cell surfaces for the agent to gain access. In other situations, the agent is not able to adapt the host's biochemistry to its own ends, for self-replication, or the environment in the host lacks essential growth factors. Active destruction of anthrax organisms apparently accounts for the natural resistance of dogs to the disease. The health of the host and environmental conditions may also make a difference to susceptibility. This is readily apparent in fish that succumb to fungal infections if their environment deteriorates. Genetic factors have an influence on susceptibility. Some of these genes have been identified, in particular the genes of the major histocompatibility complex which are involved in susceptibility to autoimmune diseases as well as some infectious disorders. There is a case for natural killer cells, as well as macrophages, being involved in natural immunity to some diseases; Chagas' disease (infection with a protozoon) is an example. Immunity to diverse diseases such as leprosy, acquired immune deficiency syndrome (AIDS), malaria, and allergies are clearly influenced by genetic background. In the plant kingdom, immunity involves a different set of mechanisms; for example, disease-resistant crops have a genetic basis. *See* ACQUIRED IMMUNE DEFICIENCY SYNDROME (AIDS); HISTOCOMPATIBILITY.

Once parasites gain entry, phagocytic cells attack them. They may engulf and destroy organisms directly, or they may need other factors such as antibody, complement, or lymphokines, secreted by lymphocytes, which enhance the ability of the phagocytes to take up antigenic material. In many cases these cells are responsible for alerting cells involved in active immunity so there is two-way communication between the innate and adaptive responses. Certain organisms, such as gonococcus, remain in the cell and, being refractory to treatment, invoke a chronic infection. Nonspecific agglutinins and lysins may also influence early development of the parasite once in the host. For obvious reasons, the innate system of host defense operates during the early phases of an immune response, and in higher organisms is inextricably associated with the adaptive immune response. *See* PHAGOCYTOSIS.

Adaptive immune response. Adaptive immunity is effected in part by lymphocytes, which are derived from bone marrow stem cells. Lymphocytes are of two types: B cells, which develop in the bone marrow or fetal liver and may mature into antibody-producing plasma cells, and T cells, which develop in the thymus. T cells have a number of functions, which include helping B cells to produce antibody, killing virus-infected cells, regulating the level of immune response, and controlling the activities of other effector cells such as macrophages.

Each lymphocyte carries a different surface receptor that can recognize a particular antigen. The antigen receptor expressed by B cells consists of membrane-bound antibody of the specificity that it will eventually secrete; B cells can recognize unmodified antigen. However, T cells recognize antigen only when parts of it are complexed with a molecule of the major histocompatibility complex. The principle of the adaptive immune response is clonal recognition: each lymphocyte recognizes only one antigenic structure, and only those cells stimulated by antigen respond. Initially, in the primary response, there are few lymphocytes with the appropriate receptor for an antigen, but these cells proliferate. If the antigen is encountered again, there will be a proportionally amplified and more rapid response. Primed lymphocytes either differentiate into immune effector cells or form an expanded pool of memory cells that respond to a secondary challenge with the same antigen.

The acquired or adaptive immune response is characterized by exquisite specificity such that even small pieces of foreign proteins can be recognized. This specificity is achieved by the receptors on T cells and B cells as well as antibodies that are secreted by activated B cells. The genes for the receptors are arranged in multiple small pieces that come together to make novel combinations, by somatic recombination. Each T or B cell makes receptors specific to a single antigen. Those cells with receptors that bind to the foreign protein and not to self tissues are selected out of a large pool of cells. For T cells, this process takes place in the thymus. The extreme diversity of T- and B-cell receptors means that an almost infinite number of antigens can be recognized. It has been calculated that potentially about 3×10^{22} different T-cell receptors are made in an individual. Even if 99% of these are eliminated because they bind to self tissues, 3×10^{20} would still be available.

Inflammation takes place to activate immune mechanisms and to eliminate thoroughly the source of infection. Of prime importance is the complement system, which consists of tens of serum proteins. A variety of cells are activated, including mast cells and macrophages. Inflammation results in local attraction of immune cells, increased blood supply, and increased vascular permeability. *See* CELLULAR IMMUNOLOGY.

Tolerance. The immune system is primed to react against foreign antigens while avoiding responses to self tissue by immunological tolerance. The neonatal immune system is particularly susceptible to tolerance induction. Although most T cells which might activate against host proteins are deleted in the thymus, these self-reactive cells are not always destroyed, and a chronic autoimmune condition results. *See* ACQUIRED IMMUNOLOGICAL TOLERANCE.

Immunization. Adaptive immunity is characterized by the ability to respond more rapidly and more intensely when encountering a pathogen for a second time, a feature known as immunological memory. This permits successful vaccination and prevents reinfection with pathogens that have been successfully repelled by an adaptive immune response.

Vaccination against smallpox was introduced by Edward Jenner in 1796, using a living agent, calf pox, that is antigenically related to smallpox. It was well known that a case of smallpox conferred life-long immunity. In fact, inhalation or dermal inoculation with smallpox, resulting in mild disease and consequent immunity, had been common in parts of Europe for decades but was not readily accepted by the medical establishment. Modern inoculation programs have been so effective that smallpox seems to have been completely eradicated.

Vaccination has been one of immunology's greatest practical successes, and mass immunization programs have led to the virtual eradication of several very serious diseases, although not always on a worldwide scale. Living attenuated vaccines against a variety of agents, including poliomyelitis, tuberculosis, yellow fever, and bubonic plague, have been used effectively. Nonliving vaccines are commonly used for prevention of bacterial diseases such as pertussis, typhoid, and cholera as well as some viral diseases such as influenza and bacterial toxins such as diphtheria and tetanus. *See* CANCER (MEDICINE); VACCINATION.

Passive immunization. Protective levels of antibody are not formed until some time after birth, and to compensate for this there is passive transfer of antibody across the placenta. Alternatively, in some animals antibody is transferred in the first milk (colostrum). Antibody may also be passively transferred artificially, for example, with a concentrated preparation of human serum gamma globulin containing antibodies against hepatitis. Protection is temporary. Horse serum is used for passive protection against snake venom. Serum from the same (homologous) species is tolerated, but heterologous serum is rapidly eliminated and may produce serum sickness. On repeated administration, a sensitized individual may experience anaphylactic shock, which in some cases is fatal. Cellular immunity can also be transferred, particularly in experimental animal situations when graft and host reactions to foreign tissue invariably occur unless strain tissue types are identical.

Hypersensitivity reactions. Reactions mediated by antibody to foreign antigens are not always beneficial to the host. Some antigen-antibody reactions produce extensive host pathology and can result in alarmingly rapid death. Such hypersensitivity reactions include anaphylaxis; serum sickness and Arthus reactions [mediated by conventional antibodies such as immunoglobulin G (IgG)]; atopic allergies, such as pollen sensitivities [mediated by immunoglobulin E (IgE)]; and various delayed reactions such as tuberculin-type reactions, or to poison ivy (a response mediated by T cells), and graft-rejection reactions (mediated by cell-bound antibodies). Toxic shock, the overstimulation of the immune system upon infection with some bacteria, is usually due to binding of their toxins to major histocompatibility complex molecules, widely and inappropriately activating T cells. *See* ANAPHYLAXIS.

In allergy, an immune response can occur to an innocuous foreign substance (allergen), such as pollen, bee stings, poison ivy, or some foods. To become allergic, an individual first needs to be exposed to the allergen and is thus sensitized. The response on reexposure is the allergic reaction. An allergy is a normal immune response, but to a substance that is not usually harmful. *See* ALLERGY.

Autoimmunity. Individuals sometimes respond immunologically to certain self antigens, so-called autoantigens, resulting in an autoimmune response. These exceptions to self tolerance are frequently associated with disease, the autoimmune diseases, which are widespread pathological conditions, including Addison's disease (adrenal insufficiency of nontuberculous origin), celiac disease, Goodpasture's syndrome, Hashimoto's thyroiditis, juvenile-onset diabetes mellitus, multiple sclerosis, myasthenia gravis, pemphigus vulgaris, rheumatoid arthritis, Sjögren's disease, and systemic lupus erythematosus. In these diseases, antibodies or T cells activate against self components.

The mechanisms responsible for autoimmune diseases are not known and may be complex. Although these diseases share some features, it is unlikely that they have a common cause. It is normally assumed that in autoimmune disease tolerance to self breaks down, eventually leading to both effector cells and antibodies against self. *See* AUTOIMMUNITY.

Suppression of immunity. In cases of unfavorable antigen-antibody reactions in autoimmunity, allergy, and transplantation, it may be desirable to suppress immune reactions. Various agents have been employed, such as x-irradiation, folate analogs, corticosteroid hormones, and antilymphocyte antisera. *See* IMMUNOSUPPRESSION. John Trowsdale

Bibliography. C. A. Janeway et al., *Immunobiology: The Immune System in Health and Disease*, 4th ed., Elsevier/Garland, 1999; J. Kuby, *Immunology*, 3d ed., Freeman, 1997; I. Roitt, J. Brostoff, and D. Male, *Immunology*, 5th ed., Mosby, 1998.

Immunoassay

An assay that quantifies antigen or antibody by immunochemical means. The antigen can be a relatively simple substance such as a drug, or a complex one such as a protein or a virus. *See* ANTIBODY; ANTIGEN.

The reactants are first mixed so that a varying quantity of one (A) is added to a constant amount of the other (B). The formation of an immune (antigen-antibody) complex is measured as a function of the varied reactant (A). The result is represented by a "standard curve" for reactant A. An unknown sample is tested by adding it to reactant B. The extent of the measured change is referred to the standard curve, and thereby is obtained the amount of reactant A which produces a comparable change. That amount is represented as the content of reactant A in the unknown sample.

This type of immunoassay requires a direct detection of only the amount of immune complex or only the amount of residual, free reactant. The immune complex can be measured by the formation of large aggregates (immunoprecipitation), of small aggregates detected by light scattering (nephelometry), or by the evolution of heat in the reaction (microcalorimetry). If one of the reactants is initially fixed to a solid support such as the wall of a test tube or a fragment of plastic, the support can be washed free of contaminants at the end of the reaction. The product of the reaction can then be assayed, for example, by immunofluorescence. *See* IMMUNOFLUORESCENCE; IMMUNONEPHELOMETRY.

Determination of residual, free reactant requires that the substance have some special property which is lost in the immune reaction. An example of this is an enzyme, such as streptococcal hemolysin O, which is inhibited by a specific antibody. The degree of inhibition can be used to quantify either the enzyme or the antibody to the enzyme.

Other forms of immunoassay require that one of the reactants (A or B) be labeled. The sensitivity of the assay, that is, its ability to detect very small amounts of a reactant, is related to the ability to detect the label. Commonly employed labels include enzymatic, fluorescent, and radioactive; others may be detected by such methods as electron spin resonance and nuclear magnetic resonance. The reactant to be measured in a test sample (AU) is of the same type as the labeled reactant (AL), but, of course, is unlabeled. AU and AL may be mixed, or AU may be added first and AL later to a constant amount of B. The reaction product of AU + AL + B now contains an amount of bound AL which varies inversely with that of added AU. By contrast, the content of free or unreacted AL varies directly with the amount of added AU. The bound AL, the free material, or both can be assayed and the results used to construct a standard curve. *See* RADIOIMMUNOASSAY.

Assays which use one reactant that adheres to a solid support again facilitate the separation of the reaction product from the unreacted material. Otherwise, the bound reactant must be separated from the free material by a subsequent procedure. That procedure is commonly precipitation with either antibody, directed against the first (specific) antibody used in the assay or substances such as ammonium sulfate or polyethyleneglycol. The label is then assayed by an appropriate procedure, for example, immunofluorescence or enzyme-linked immunoassay.

Enzymatic labels can achieve great sensitivity, similar to that of radioimmunoassay. They can be used in two ways: (1) The enzyme is initially reacted with a low-molecular-weight substrate that is to be quantified. The antibody to this substance inhibits the action of the enzyme. The addition of a free substance reverses the inhibition of the enzyme by the antibody. The extent of reversal of the inhibition, or of the residual activity of the enzyme, is related to the quantity of the measured substance. (2) The enzyme is coupled to a macromolecular reactant. The activity of the enzyme then reflects the concentration of the labeled reactant. This, in turn, is related to the concentration of the unlabeled reactant in the test sample.

The enzymatic label can be measured by various assays. The simplest are colorimetric, which use a color change. Other changes can be measured by absorptiometry, chemiluminescence, fluorometry, nephelometry, potentiometry, or radioactive assay. *See* IMMUNOLOGY. Alexander Baumgarten

Bibliography. J. W. Kimball, *Introduction to Immunology*, 3d ed., 1990; W. E. Paul, *Fundamental Immunology*, 4th ed., 1998; D. M. Weir (ed.), *Handbook of Experimental Immunology*, 5th ed., 1996; D. Wild (ed.), *Immunoassay Handbook*, 2d ed., 2000.

Immunochemistry

A discipline concerned both with the structure of antibody (immunoglobulin) molecules and with their ability to bind an apparently limitless number of diverse chemical structures (antigens); with the structure, organization, and rearrangement of the genes coding for the immunoglobulin molecules; and with the structure and function of molecules on the surface of animal cells, such as the transplantation (histocompatability) antigens, which recognize antibodies and the thymus-derived lymphocytes mediating the cellular immune response. A development of immunochemistry that is of medical, scientific, and commercial importance uses the binding specificity of immunoglobulins to measure complex (but sometimes simple) chemical structures such as alkaloids, hormones, proteins, peptides, complex carbohydrates, and lipids. The tests used for such measurements employ the binding of radioactive antigens (the radioimmunoassay) or the ligation of an antigen bound to an enzyme (the enzyme-linked immunoassay). These assay methods are highly sensitive and specific.

The impact of immunochemistry has been very strong on medicine, giving pathophysiological insights into the mechanism of diseases such as multiple myeloma, chronic infections, and various disorders of immune complex formation. It seems likely that precise delineation of cell recognition and cell-cell interactions may yield information which will make easier the transplantation of organs between incompatible donors and recipients, and will also aid the elimination of aberrations of the immune system which at present cause diseases in humans. The applications of immunoassays for measuring contaminants in industry and for the detection of toxic agents in the environment is also likely to expand considerably. *See* CELLULAR IMMUNOLOGY; IMMUNOASSAY; IMMUNOLOGY; RADIOIMMUNOASSAY; TRANSPLANTATION BIOLOGY. W. H. Konigsberg; F. F. Richards

Bibliography. L. E. Glynn and M. D. Steward (eds.), *Immunochemistry: An Advanced Textbook*, 1977; J. W. Kimball, *Introduction to Immunology*, 3d ed., 1990; C. J. Van Oss and M. H. V. Van Regenmortel (eds.), *Immunochemistry*, 1994.

Immunoelectrophoresis

A combination of the techniques of electrophoresis and immunodiffusion used to separate the components of a mixture of antigens and make them visible by reaction with specific antibodies.

Most biological substances in solution carry either an excess of protons (+1, +2, etc.) or a deficiency (−1, −2, etc.) and therefore can be separated according to their charge in an electrical field by electrophoresis. Substances which carry a similar net charge tend to occupy a similar position after separation. These substances, however, are generally antigenic and can react with antibodies to form immunoprecipitates which, in clear media (such as agar or gelatin), appear as opaque, white lines because they scatter light, whereas the media themselves transmit light without hindrance. The latter technique is known as immunodiffusion. *See* ELECTROPHORESIS; PRECIPITIN TEST.

Method. A medium such as agar is deposited on a convenient base, for example, a microscopic slide. A small well (0.04–0.20 in. or 1–5 mm in diameter) is cut in the medium. A test solution is deposited in the well, and the contained substances are separated by electrophoresis along one axis of the plate. A trough is then cut in the medium parallel to, but at some distance (0.12–0.20 in. or 3–5 mm, for example) from, the line of the separated substances. The trough is filled with antiserum which contains antibodies to one or more of the separated substances. The antiserum and substances diffuse toward one another and, where they meet, form curvilinear patterns of precipitation. These patterns can be seen directly in clear media or can be visualized after washing out unreacted materials and staining in opaque media.

Variants. Many variants of the technique have been developed.

Crossed electrophoresis. This method uses an initial electrophoretic separation along one axis of the plate. After separation, a strip of medium along the axis is preserved, while the remainder is excised. New medium, which contains antiserum, is then poured beside the central strip. The plate is turned 90°, and electrophoresis is resumed. Lines of precipitation form as the initially separated substances (antigens) are driven by their electrical charge out of the central strip and into the antibody-containing medium.

"Rocket" electrophoresis. This is a variant of crossed electrophoresis in which the medium typically contains only one antibody. No initial electrophoresis is done, but the test substances are driven directly into the medium, which contains the antibody. "Λ"- or "rocket"-shaped trails of precipitation appear within suitable ratios of concentration of the reactants. The height of the rocket is proportional to the concentration of the test substances and can be used to quantify as little as 0.1 μg/ml of protein.

Counterimmunoelectrophoresis. This technique uses two wells or points of application, one above the other, along the electrical axis. The anodal well is filled with antibody, and the cathodal with a negatively charged antigen. Electrophoresis results in the antibody migrating cathodally, while the antigen moves anodally. A line of precipitation appears where the two meet. This technique has allowed rapid detection of substances, for example, those of infectious origin.

Further variants. Other variants of the above techniques can be subdivided into those which: (1) substitute another method of separation for the initial electrophoresis, such as isoelectric focusing, gel filtration, or chromatography; (2) enhance resolution by the addition before separation of a marked tracer antigen or antibody, such as a radioactively labeled one, followed by autoradiography; or (3) use a vertical overlay of the plate with, or with incubation in, an antibody with or without a label (radioactive, enzymatic, and so on) to enhance visibility after separation. In immunofixation the initial separation (which may be electrophoretic) is followed by an overlay with antibody, instead of its application to a trough at the side. The separated components are precipitated in place, thus helping to identify variously charged forms of the same antigen prior to their undue diffusion.

Immunoelectrophoresis or its variants can also be combined with methods which exploit an inherent property of the antigen to enhance visibility and resolve constituents that possess the selected property. For instance, the precipitates can be stained with dyes that react only with lipid or carbohydrate constituents, or the antigens in the precipitates can be identified by their enzymatic properties. *See* IMMUNOASSAY. Alexander Baumgarten

Bibliography. O. G. Bier et al., *Fundamentals of Immunology*, 1986; A. J. Crowle, *Immunodiffusion*, 1973; J. J. Scheidegger, *Int. Arch. Allergy Appl. Immunol.*, 7:103, 1955.

Immunofluorescence

A technique that uses a fluorochrome to indicate the occurrence of a specific antigen-antibody reaction. The fluorochrome labels either an antigen or an antibody. The labeled reactant is then used to detect the presence of the unlabeled reactant, for example, a component of a section of body tissue seen under a microscope. The use of a labeled reactant (such as an antibody which both detects and indicates the antigen) to reveal the presence of an unlabeled one is termed direct immunofluorescence. The use of a labeled indicator antibody, which reacts with an unlabeled detector antibody that has previously reacted with an antigen, is termed indirect immunofluorescence. Both techniques can be used qualitatively to indicate the presence of a specific reactant in a preparation. They can also be used semiquantitatively, for example, to determine the highest dilution (lowest concentration) of a tested detector reactant that can detect a given substance. Substitution of a light meter for the human eye permits a quantitative measurement in immunofluorometry. *See* FLUORESCENCE.

Background fluorescence. Many tissues and substances of biological origin are capable of autofluorescence. The emitted light is often blue or blue-green, but may be another color. Autofluorescence contributes to the background fluorescence of the examined material. Another contributor to the background is the residual deposit of the indicator material which cannot be removed by washing but is retained by nonspecific adherence to the examined material. The specific fluorescence must increase above the background in order to be perceived. This can be obtained in several ways. For instance, the indicator may be a substance which emits light of a different wavelength than the natural autofluorescence of the tested material, or the indicator may concentrate over accumulations of the tested reactant so that its brightness there is much greater than the general level of background fluorescence.

Efficiency of detection. Direct immunofluorescence generally produces low background, but also a low level of specific illumination. Indirect immunofluorescence tends to increase both the background and specific illumination, but because several molecules of the detector can react with each molecule of the tested reactant and, in turn, several molecules of the indicator can react with each molecule of the detector, a marked enhancement occurs in the efficiency of detection. Therefore, indirect immunofluorescence can detect less reactant than direct. Furthermore, the indirect technique is more convenient because one indicator antibody (such as goat antirabbit immunoglobulin) can be used with various detector antibodies (such as those of rabbit origin). The successive application of antibodies in the indirect technique produces a "sandwich" of which only the last layer fluoresces. Theoretically, the layers in the sandwich could be increased indefinitely, with a concomitant augmentation in sensitivity. In practice, each layer deposits some nonspecifically bound material, to which the next layer reacts specifically, as well as adding its own nonspecific load. The background therefore becomes excessively great within three or four layers.

Fluorochromes. The most common fluorochromes used for immunofluorescence are fluorescein isothiocyanate (with maximal absorption at 495 nanometers), which emits yellow-green light, and tetramethylrhodamine isothiocyanate (with maximum absorption at 544 nm), which emits red light. Both react spontaneously with the amino groups of proteins in alkaline solution. Other fluorochromes include DANS (1-dimethylaminonaphthalene-5-sulfonic acid) and RB200 (lissamine rhodamine B) sulfonyl chlorides. Two fluorochromes, each marking a different antibody, can be used for simultaneous identification of two antigens. Separate localization of each antigen is visualized by one color of fluorescence (such as green or red), but congruence of the antigens produces a mixed color (such as orange).

Light source. Observation of immunofluorescence by transmitted light requires an exciting beam of light below the specimen and a barrier filter above to absorb the exciting but not the emitted light. Alternatively, the exciting light is refracted down the objective of a microscope and passed through the specimen. The emitted light, however, travels back up the objective to the eyepiece (incident light fluorescence). Excitation is usually provided by ultraviolet light sources, such as a mercury lamp, a xenon arc, or halogen-filled lamps. It should be noted that sources of ultraviolet light can damage unprotected eyes.

Adaptations. Novel adaptations of immunofluorescence include immunoenzymatic assays and quenching. The former use a reactant coupled to an enzyme (β-galactosidase, for example) which cleaves a nonfluorescent compound (such as a derivative of umbelliferone) to yield a fluorescent derivative (umbelliferone, for example). Quenching uses, say, two fluorochromes, one of which absorbs light emitted by the other. One fluorochrome labels the antigen, another the antibody. Antigen-antibody complexes retain both fluorochromes, so that the initially emitted light is absorbed and so quenched by the second compound. The extent of quenching is assayed. *See* IMMUNOASSAY. Alexander Baumgarten

Bibliography. C. C. Curtain and A. Baumgarten, *Immunology*, 10:499, 1966; J. W. Kimball, *Introduction to Immunology*, 3d ed., 1990; G. Wick, K. N. Traill, and K. Schaunstein (eds.), *Immunofluorescence Technology*, 1983.

Immunogenetics

A scientific discipline that uses immunological methods to study the inheritance of traits. Traditionally, immunogenetics has been concerned with moieties that elicit immune response, that is, with antigens (antigenic determinants). It has now broadened its scope to study also the genetic control of the individual's ability to respond to an antigen.

Methods. The immunological methods used in immunogenetics are of two principal kinds, serological and histogenetical. In serological methods, antibodies are used to detect antigens, either in solution or on a cell surface. In histogenetical methods, immune cells (lymphocytes) are used to detect antigens on the surface of other cells. *See* ANTIBODY; ANTIGEN.

Serological methods. To produce antibodies or to stimulate immune cells, an antigen is introduced into the blood of an animal—usually a mammal, but other vertebrates that have developed an immune system can also be used. Blood delivers the antigen to the lymphoid organs, where it is picked up by specialized cells which then present it to B and T lymphocytes. The B lymphocytes react to the stimulation by secreting antibodies (immunoglobulins) into the blood plasma; T lymphocytes react by differentiating into effector cells. The antibodies are then collected by bleeding the animal and separating the serum from the clotted blood. Serum containing antibodies specific for a given antigen is called immune serum or antiserum. The activated B lymphocytes can also be explanted from the animal and immortalized by fusion with an appropriate tumor cell. The resulting hybridomas then continuously produce antibodies of one kind (such antibodies are called monoclonal because they derive from a single lymphocyte clone). Monoclonal antibodies can be collected from the supernatant of a tissue culture in which the hybridoma is growing, or collected from the ascites fluid filling the peritoneal cavity of an animal into which the hybridoma has been inoculated. *See* CELLULAR IMMUNOLOGY; MONOCLONAL ANTIBODIES; SERUM.

The antibody binds only to the antigen that elicited its secretion, or at most to a few related compounds; it does not bind to any other antigens. There are five principal ways of measuring this antigen-antibody binding: (1) In the precipitation test, the soluble antigen and soluble antibody form insoluble complexes that fall out as visible or visualizable precipitate from a solution. (2) In the complement-fixation test, the antigen-antibody complexes bind (fix) complement, and the depletion of complement is then determined with the help of a test system. (3) In the hemolytic and cytotoxic tests, antibody binds to antigen on a cell surface and the complex activates complement components, some of which then kill erythrocytes or other cells. (4) In the agglutination test, the antibodies bind with one site to antigen on one cell and with the other site to antigen on another cell and thus glue the cells together. (5) In the fluorescence test and the radioimmunoassay, a label is attached to the antibody (fluorescein dye or radioactive atoms), which then signals to the investigator that the antibody has remained attached to the cell surface. *See* AGGLUTINATION REACTION; COMPLEMENT-FIXATION TEST; IMMUNOFLUORESCENCE; PRECIPITIN TEST; RADIOIMMUNOASSAY.

Histogenetical methods. The reaction of the immune lymphocytes with the target antigens can be revealed by three principal assays. In one, graft rejection methods, the lymphocytes trigger a reaction that destroys a foreign tissue or organ transplanted to a recipient. In the second assay, the immune lymphocytes proliferate in response to an antigen: this is called the mixed lymphocyte reaction if the antigen is present in the membrane of a foreign lymphocyte; it is called the T-lymphocyte proliferation assay when a soluble antigen is picked up and presented by a specific cell, which is usually genetically identical with the lymphocyte. In the third assay, the cell-mediated lympholysis assay, the activated lymphocytes differentiate into cytolytic cells which kill appropriate target cells. *See* TRANSPLANTATION BIOLOGY.

Genetic situations. There are four genetic situations in which immune cells and antibodies can be elicited. In the xenogeneic situation, the antigen derives from a different animal species than the responding animal. In the allogeneic situation, the donor is one individual and the responder another, genetically disparate individual of the same species. In the syngeneic situation, the donor and the recipient are genetically the same (for example two individuals of an inbred strain). In the autologous situation, an individual responds to its own antigens. Responses in syngeneic and autologous situations occur only under pathological conditions when the individual fails to distinguish between self and nonself entities and regards its own molecules as foreign antigens (autoimmunity). *See* AUTOIMMUNITY.

Molecular methods. In modern immunogenetics research, the serological and histogenetical methods are combined with molecular methods in which the researcher isolates and works with the genes that code for the traits. Usually, the antigens are identified with antibodies or lymphocytes, then messenger ribonucleic acid (mRNA) coding for the antigen-bearing protein is isolated and used to identify the deoxyribonucleic acid (DNA) segment (gene) from which it has been transcribed. The isolated gene, in turn, can be inserted into a suitable cell and activated, and the expressed molecules used to elicit new sets of antibodies or immune cells for a refined analysis of the gene product. This approach of going back and forth from classical to molecular methods has proved to be very successful and has led to the elucidation of several complex genetic systems.

Tools. Animal immunogenetics relies heavily on the use of inbred, congenic, and recombinant inbred strains.

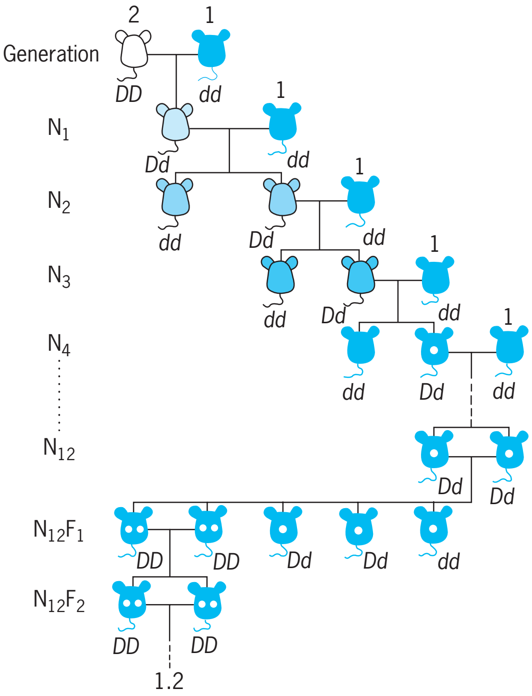
Inbred strains. Inbred lines result when individuals that are more closely related to each other than randomly chosen individuals mate together, for many generations. In a randomly mating population, each individual is homozygous for most of its genes (carries identical alleles at each of the loci) but heterozygous at some 20 to 40% of loci. Random mating propagates this status, whereas inbreeding tends to increase the proportions of homozygous loci.

The reason for this increase is easy to understand. First, assume that the two parents of an initial mating are both *Aa* heterozygotes (where *A* and *a* are alleles at a given locus). The *Aa* × *Aa* mating produces *AA*, *Aa*, and *aa* progeny. If brothers are mated with sisters from this litter, sooner or later two *AA* or two *aa* individuals will be selected by chance and mated together. When this happens, all the progeny in all subsequent brother × sister matings will always be *AA* or *aa*. Homozygosity has been achieved at the *A* locus, and as long as the individuals in this genealogical line are not outcrossed to unrelated individuals, the line remains homozygous for this locus. The same thing happens sooner or later for other loci, *B*, *C*, *D*, and so on—for all those loci that were originally in a heterozygous state. The mathematical theory of inbreeding predicts that after 20 generations of brother × sister matings, more than 97% of the loci become homozygous. By convention, this figure is considered sufficient for a strain to qualify as inbred.

When the matings are not between brothers and sisters but between some more distant relatives, it takes proportionally longer to achieve this degree of homozygosity. Because each individual carries a certain load of deleterious genes (that is, genes that in a homozygous state lower the viability of their bearer), and because inbreeding also increases the homozygosity of these genes, an inbreeding depression may set in after a few inbreeding generations. In this phase of the inbreeding program, many individuals either die or do not breed. However, if the program has been initiated on a sufficiently large scale, there are usually enough genetically healthy individuals to carry the program over this hurdle. After 20 generations of brother × sister mating, individuals within a group become genetically homogeneous, resembling one another the way identical twins do; an inbred strain (line) is thus established. A number of inbred strains have been developed in the laboratory mouse, rat, domestic fowl (chicken), guinea pig, Syrian hamster, and rabbit.

Congenic lines. The advantage in working with inbred strains rather than outbred animals is that inbred strains restrict the variability of the conditions of an experiment. However, when two strains are compared and it is found that they respond differently to a treatment, it is not known to what gene this difference should be attributed. The strains may differ at as many genetic loci as two unrelated individuals in an outbred population do. To study the effect of single, defined genes, immunogeneticists have developed congenic lines. These lines always come in groups, the smallest group being a pair, which consists of a congenic line and its inbred partner strain. The two are homozygous at more than 97% of their loci (that is, they are inbred) and are identical except, ideally, at one locus—the locus that is to be studied. At this differential locus, the two strains carry different alleles. This ideal situation is achieved only rarely, if ever; most of the time immunogeneticists are satisfied when the two strains in a pair differ in a short chromosomal segment that includes the differential locus. In addition to pairs, trios, quartets, or quintets of congenic lines can also be produced, in which all the lines differ from the inbred partner strain and from one another in that they carry different alleles at the same differential locus.

The simplest way of producing a pair of congenic lines involves two strains, 1 and 2, that differ at locus *D* (see **illus.**). Strain 1 is a *dd* homozygote and strain 2 is a *DD* homozygote. To study the effect of the *D* gene, a line is produced that is genetically identical with strain 1 but carries the *D* allele. To this end, the two strains are crossed, and the resulting F_1 hybrid is



A scheme of congenic-line production by repeated backcrossing to strain 1 and selection in each segregating generation for the *D* allele donated by strain 2. The successive generations of matings of *Dd* individuals with strain H-1 are designated N_1, N_2, N_3, and so on. After 12 such generations, two *Dd* individuals are intercrossed and *dd* individuals are selected and mated among themselves. The intercrossing generations are then designated F_1, F_2, F_3, and so on. (*After J. Klein, Immunology, John Wiley and Sons, 1982*)

backcrossed to strain 1. In the backcross generation, the *D* gene segregates so that half the progeny is *Dd* and the other half *dd*. The animals are then typed and one of the *Dd* individuals is selected for further mating (backcrossing) to strain 1; this procedure is repeated for several generations. At the end, two *Dd* individuals are intercrossed and *DD* animals are obtained. By brother × sister mating of *DD* animals, a 1.2 line is obtained which is very similar to strain 1, but which carries the *D* instead of the *d* allele at the *D* locus. According to calculations, 99% of the loci not linked to *D* (that is, on a different chromosome than *D*) are of strain 1 origin when the backcross is repeated 12 times.

The largest number of congenic lines is available in the laboratory mouse. Lines which are known to differ from the inbred partner strain at a single locus (rather than a chromosomal segment) are called coisogenic. They arise when one of the loci in an inbred strain mutates and thus changes to another allele. Individuals homozygous for the mutation are then coisogenic with the original strain.

Recombinant inbred strains. To find out whether two loci are on the same or on different chromosomes, two individuals that differ in the traits controlled by these loci are mated and then the F_1 hybrids are intercrossed. In the F_2 generation that results from this intercross, the genes assort either independently, if they are on different chromosomes, or nonrandomly, if they are on the same chromosome—that is, when they are linked. Each time the strains are tested for linkage, this laborious procedure must be repeated. To avoid this repetition, immunogeneticists have prepared a "frozen" F_2 generation by establishing separate inbred lines from the different F_2 individuals. Such lines are called the recombinant inbred strains. Similarity of recombinant inbred strain-distribution patterns of two traits suggests that they are controlled by loci on the same chromosome. A number of recombinant inbred strains are available in the laboratory mouse—one of the most favored experimental animals.

Scope. Contemporary immunogenetic research concentrates on two main categories of antigenic substances—those present in body fluids, primarily blood serum or plasma, and those expressed on surfaces of various cells.

Body-fluid antigens. In this category, a prominent position is occupied by immunoglobulins (Ig), the proteins that constitute antibodies. Although antibodies are usually used to detect antigens, they themselves may also serve as antigens, and antibodies can be produced against them. These antibodies against antibodies detect three principal kinds of antigenic determinants: isotypic, allotypic, and idiotypic.

1. *Isotypic determinants*. These are shared by all individuals of a given species but absent in individuals of other species. Antibodies against isotypic determinants can normally be produced by xenogeneic immunization, that is, by injecting immunoglobulins of one animal species into individuals of another species. Isotypic determinants differentiate classes and subclasses of immunoglobulins. There are five classes of human immunoglobulins: IgG, IgM, IgA, IgD, and IgE. Some of the classes (IgG, IgA) consist of several subclasses (IgG 1–4; IgA 1, 2) which are more related to one another than two different classes are. Both classes and subclasses differ from one another not only in their serological but also in their physiochemical properties. *See* IMMUNOGLOBULIN.

2. *Allotypic determinants (allotypes)*. These are determinants present in some individuals of a given species and absent in other individuals. Antibodies against allotypic determinants arise normally during allogeneic immunizations; abnormally, they are present in the sera of patients with certain diseases, in particular, rheumatoid arthritis. In the latter case, the antibodies react with self components: they are autoantibodies. Most identified allotypic determinants are present in the constant region of the Ig molecule, but a few have also been located in the variable region. In many instances, allotypic differences have been correlated with amino acid substitutions in the primary structure of the Ig polypeptide chain. One or two substitutions are usually enough to create an allotypic difference between two Ig chains. Allotypic determinants have been studied extensively in human, rabbit, and mouse. The best known of the human allotypes is the Gm system, expressed on the different subclasses of the IgG class. A number of Gm determinants have been identified and their distribution in human populations mapped.

3. Idiotypic determinants (idiotypes). These are unique antigenic determinants restricted to antibodies of a single species or of only a few related species. They can be identified with antibodies produced during xenogeneic, allogeneic, or syngeneic (autologous) immunization. They reside in the variable region of the Ig molecule on either the heavy or the light chain. Differences in idiotypic determinants arise because individuals of the same species differ in the set of genes controlling the variable-region polypeptide chains, and because in ontogeny of the B lymphocyte these genes undergo rearrangements that change the structure of the encoded proteins. Because of their location in the V region, idiotypic determinants are associated with antibody specificity in that Ig molecules reacting with different antigens usually express different idiotypes. *See* IMMUNOLOGICAL ONTOGENY.

Cell-surface antigens. The main categories of cell-surface molecules studied by immunogenetical methods are blood group antigens, histocompatibility antigens, tissue-restricted antigens, and receptors.

1. *Blood-group antigens*. These are alloantigens found on erythrocytes. Historically, they were the first antigens studied by immunogenetical methods, and a great number of them have been identified in both humans and animals. The blood-group variations usually reflect differences in the structure of cell-surface carbohydrates. *See* BLOOD GROUPS.

2. *Histocompatibility (H) antigens*. These are antigens capable of inducing cellular immune responses and hence are detectable by histogenetical methods. However, at least some of them also stimulate antibody responses. They are divided into two

groups, minor and major, but the two groups have nothing else in common than the fact that they are detected by the same method.

The minor *H* loci are quite numerous (there may be several hundred of them), and they are scattered all over the genome. They are referred to as minor because they act as weak antigens in inducing cellular immune response. However, they act synergistically and their combined effect can be quite strong. Their function is not known.

In contrast, most of the major *H* loci reside in a single chromosomal region and are potent inducers of cellular immune response. The function of the major histocompatibility complex (MHC) is to introduce antigens to the T lymphocyte. The T lymphocyte does not recognize antigens if they are not accompanied by MHC molecules on the antigen-presenting cell. The MHC of different species has different names. In humans it is called the human leukocyte antigen (HLA) complex.

The MHC genes can be divided into two classes, which differ in their organization and function. Class I and II genes comprise polypeptide chains that are folded in such a way that a groove is created in the external portion of the molecule. During the synthesis of the MHC molecule, the groove binds short peptides derived from other proteins of the cell. The molecule then displays the peptides on the cell surface. Peptides derived from the body's own proteins are ignored by the immune system; those derived from invading parasites are recognized (together with portions of the MHC molecule itself) by T lymphocytes, the recognition activates the lymphocytes, and the activated lymphocytes initiate the specific form of the immune response. The class I MHC molecules bind peptides derived from proteins manufactured in the cell (either the cell's own proteins or those of intracellular parasites). The bound foreign peptides are recognized by cytotoxic T lymphocytes, whose function in the immune response is to kill other cells. The class II MHC molecules, however, bind peptides derived from proteins that have been taken up by the cell from the extracellular milieu. The bound foreign peptides are recognized by helper T lymphocytes, which regulate the immune response and facilitate the production of antibodies by B lymphocytes.

The functional class I and class II genes are highly polymorphic, which means that they occur in a number of allelic forms carried by different individuals. This is apparently unique and characteristic of the MHC. Its significance is unclear, but the polymorphism may represent a way of coping with MHC-associated nonresponsiveness to certain antigens. Certain combinations of antigen and MHC molecules are not recognized by T lymphocytes, so that an individual carrying the particular MHC molecules (alleles at a particular MHC locus) is unresponsive to the particular antigen. Because the nonresponsiveness depends on the presence of a particular MHC allele, when responders mate with nonresponders, the ability to respond segregates in the progeny with the MHC. Outwardly it appears as if there were special immune response (*Ir*) genes linked to the MHC, while, in fact, the MHC themselves are the *Ir* genes. Because certain MHC molecules confer nonresponsiveness to certain antigens, they may make certain individuals vulnerable to certain pathogens. The MHC polymorphism may be the consequence of selection for MHC alleles that allowed individuals to survive past epidemics. *See* POLYMORPHISM (GENETICS).

In the human population, certain HLA alleles or combinations of alleles (haplotypes) are associated with disease. Individuals suffering from a specific disease often possess a particular HLA allele more frequently than healthy individuals. For example, nearly all individuals suffering from narcolepsy (a disease characterized by a sudden, uncontrollable compulsion to sleep during waking hours) have the *HLA-DRB1***02* allele. *See* HISTOCOMPATIBILITY.

3. *Tissue-restricted antigens*. These are expressed on some tissues but not on others and therefore serve as markers for cell sorting. A number of such antigens have been identified in humans and the mouse, particularly on blood cells (many blood-group antigens fall into this category). There are antigens that are expressed on T but not B lymphocytes and others for which the opposite is true. Some antigens are expressed only on a subpopulation of T or B lymphocytes.

4. *Receptors*. These are molecules that are capable of specifically interacting with certain other molecules. The interaction often leads to activation or inhibition of the receptor-bearing cell. The receptors themselves can serve as antigens and can elicit antibody production or activation of immune lymphocytes. The most extensively studied receptors are those that recognize antigens and are located on the surface of B and T lymphocytes. The B-lymphocyte receptor closely resembles the antibody molecules that are secreted by the cell upon activation and differentiation. The T-lymphocyte receptors are related to but different from the immunoglobulin molecules. *See* GENETICS; IMMUNOLOGY.

Jan Klein

Bibliography. J. Klein, *Immunology*, 2d ed., 1997; J. Klein, *Natural History of the Major Histocompatibility Complex*, 1986; R. Srivastava, G. P. Ram, and P. Tyle (eds.), *Immunogenetics of the Major Histocompatability Complex*, 1991.

Immunoglobulin

Any of the glycoproteins in the blood serum that are induced in response to invasion by foreign antigens and that protect the host by eradicating pathogens. Antibodies belong to this group of proteins. An antigen is any substance capable of inducing an immune response. Intact antigens are able to specifically interact with the induced immunoglobulins. Normally, the immune system operates in a state known as self-tolerance, and does not attack the host's own tissues, but occasionally the immune system targets host-specific antigens, resulting in autoimmune disease. *See* AUTOIMMUNITY.

Immunoglobulins are composed of two identical heavy (H) and two identical light (L) polypeptide chains. Each H and L chain has an amino-terminal variable (V) region and a carboxyl-terminal constant (C) region. Although V regions from different antibodies exhibit considerable sequence variation, there is a large degree of sequence similarity among C regions of different antibodies.

In the living animal, antibodies first bind to antigen and then, ideally, eliminate it as a threat to the host. The antigen is bound to the antibody at the antigen-combining site, a cleft or groove formed by the association of the amino-terminal variable regions of both chains, which are complementary to the size, shape, and charge characteristics of the antigen. Bound antigen is then inactivated by one or more of the effector functions of the antibody.

Human myeloma proteins, the first structurally discrete immunoglobulins known, result from the malignant expansion of a normal clone of antibody-producing cells. The antigenic specificity of myeloma proteins, however, is generally unknown.

Structure. An immunoglobulin molecule has a Y-shaped appearance with two arm segments which are the Fab (fragment antigen binding) region connected by a flexible hinge to a stem composed of the Fc (fragment crystallizable) region. The basic structural unit of immunoglobulins is the immunoglobulin domain. The L chain is composed of two domains, and the H chains are divided into four domains for IgG, IgA, and IgD and a fifth domain for both IgM and IgE. Each domain contains approximately 110 amino acids and is encoded by a separate deoxyribonucleic acid (DNA) exon. *See* EXON.

Classes of immunoglobulins. Immunoglobulins are heterogeneous with respect to charge, size, antigenicity, and function. There are three categories of antigenic determinants present on immunoglobulins: isotypes are found in all individuals, allotypes are found in some individuals, and idiotypes are associated with the amino-terminal variable region. Isotypic determinants are located on the carboxyl-terminal constant region and are used to group immunoglobulin H and L chains into isotypes or classes. In total, there are five human H-chain classes. IgM contains mu (μ) H chains, IgG contains gamma (γ) H chains, IgA contains alpha (α) H chains, IgD contains delta (δ) H chains, and IgE contains epsilon (ϵ) H chains. IgG has four subclasses, IgG1, IgG2, IgG3, and IgG4, while IgA has two subclasses, IgA1 and IgA2. There are two L-chain isotypes named kappa (κ) and lambda (λ). Kappa and lambda chains may be associated with H chains of any isotype, and a complete description of an immunoglobulin molecule requires identification of both H and L chains.

IgG. IgG is the most abundant immunoglobulin class in the serum. The four IgG subclasses, based on chemical distinctions between the H chains, are named numerically in order of their relative concentration in normal serum. All four subclasses contain approximately 3% carbohydrate. However, the main structural difference among them is in the hinge region, where the number of amino acids and the number of interchain disulfide bonds vary. IgG isotypes are associated with complement fixation, opsonization (that is, rendering more susceptible to phagocytosis), fixation to macrophages, and membrane transport. All four IgG subclasses bind streptococcal protein G. IgG1, IgG2, and IgG4 isotypes bind protein A from *Staphylococcus aureus*. *See* STAPHYLOCOCCUS; STREPTOCOCCUS.

IgA. Of the two IgA subclasses, IgA1 is the predominant subclass of IgA in human serum, representing 7–15% of serum immunoglobulin. IgA1 is the dominant subclass in all external secretions, including milk, saliva, tears, and bronchial fluids. The percentage of subclass IgA2 is higher in these fluids than in serum. Secretory IgA consists of two IgA monomers, one J chain, and one glycoprotein called secretory component. In IgA, J chain may be necessary for polymerization of the two monomeric units; however, it is not required for the polymerization or secretion of IgM antibodies. The secretory component is cotransported with the antibody. Its function is unknown, but possibly it stabilizes the IgA molecule and provides protection from the enzymatic activity of the gastrointestinal tract.

IgM. IgM is the first immunoglobulin to appear during the primary immune response. IgM is also a minor component of secretory immunoglobulins at mucosal surfaces and in breast milk. IgM antibodies are secreted from B lymphocyte cells as pentameters or hexameters.

IgD and IgE. IgD and IgE are present in minute amounts in normal human serum. No function has been clearly attributed to IgD. IgE is active against parasites and acts as a mediator of immediate hypersensitivity. Circulating IgE binds the high-affinity receptor on the surface of basophils in the circulation and mast cells in tissues. Subsequent to the interaction of bound IgE with antigen, basophils disintegrate, releasing inflammatory agents such as histamine. This type of response leads to allergy and anaphylaxis (hypersensitivity following injection of an antigen) in susceptible individuals. Atopic dermatitis, an inflammatory skin disorder of unknown etiology, is characterized by excessive production of IgE. *See* ANAPHYLAXIS; ANTIBODY; ANTIGEN; ANTIGEN-ANTIBODY REACTION; IMMUNOLOGY; PROTEIN.

Kathleen N. Potter; J. Donald Capra

Bibliography. W. E. Paul (ed.), *Fundamental Immunology*, 4th ed., 1998; F. Shakib, *Molecular Aspects of Immunoglobulin Subclasses*, 1990; G. P. Smith, *The Variation and Adaptive Expression of Antibodies*, 1990.

Immunoglobulin genes

The genes encoding the genetic information to make antibodies (immunoglobulins) of different specificities and classes. Specificity is encoded in the variable region of the antibody, which is generated by assembling gene segments through a process called rearrangement. Biological activity of an antibody is established through a process called class switching

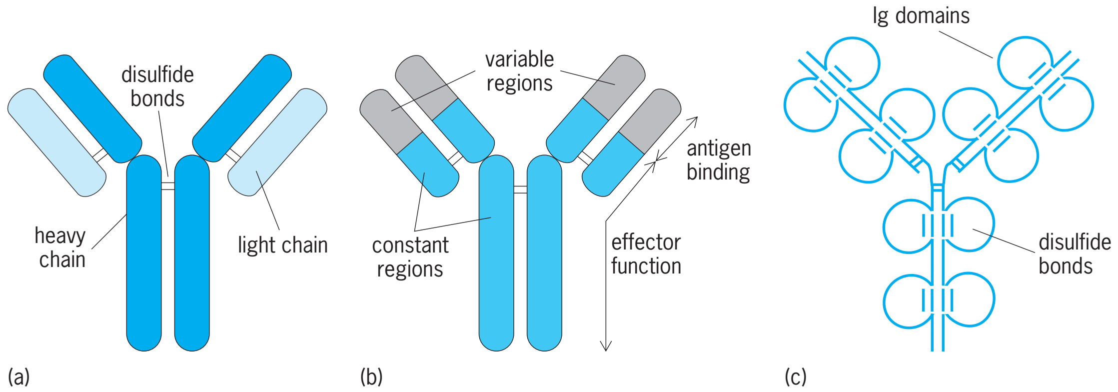


Fig. 1. Structure of a generic antibody (IgG). (*a*) A heavy chain is connected to a light chain by a disulfide bond. This heterodimer is linked to another heavy/light-chain heterodimer by another disulfide bond. (*b*) Amino acid sequencing showed two regions within one chain, a constant region conveying the biological activity (effector function) of the antibody and a variable region that specifies the antigen binding. (*c*) More detailed amino acid sequencing identified Ig domains, shown here as loops that are created by disulfide bonds. Variable regions consist of one domain; constant regions of the heavy chain have three or four Ig domains, whereas light chain constant regions have only one.

in the constant part of the antibody. *See* ANTIBODY; GENE; IMMUNOGENETICS; IMMUNOGLOBULIN.

Structure. Once researchers observed that mice injected with mineral oil would develop B cell (bone marrow–derived lymphocyte) tumors that secreted immunoglobulins (antibodies, Ab) that appeared to be all the same (monoclonal), the groundwork was laid for solving the structure of antibodies. These uniform immunoglobulins (which are glycoproteins) could be purified and studied to determine the order of amino acids in the protein or, by treating the protein with chemicals or enzymes, to deduce the structure of the protein of interest.

Rodney Porter and Gerald Edelman chose the latter approach and solved the structure of immunoglobulins by treating the immunoglobulin (Ig) with a reducing agent (β-mercaptoethanol). They found that immunoglobulins consist of two identical small protein chains and two identical large protein chains. These are generally referred to as the light and heavy chains of an antibody, respectively. To form the final structure, one light chain pairs with one heavy chain, they are linked together by disulfide bonds (S-S bonds between two amino acids), and these heterodimers (larger protein structures made of two different protein chains) are then paired with another heterodimer and held together also by disulfide bonds making a final 4-chain structure (**Fig. 1***a*).

Amino acid sequencing also revealed extreme heterogeneity in the first (amino-terminal) 110–125 amino acids of the heavy and light chains. Investigators concluded that this was the part responsible for the unique antigenic specificity of an Ab. These areas were called variable (V) regions (Fig. 1*b*). The two variable regions of the heavy and light chains together create the binding site of the antibody. This binding site recognizes the antigen in a very specific way, similar to the way one specific key fits one particular lock. *See* ANTIGEN.

The more conserved part of the sequence was named the constant (C) region (Fig. 1*b*). The C region of the heavy chain in particular mediates the biological activity or effector function of the antibody.

The amino acid sequence also revealed that the variable and the constant regions shared a common domain structure. These Ig domains (Fig. 1*c*) were roughly 110 amino acids long and showed some level of sequence and structure conservation, such as a disulfide bond creating a loop of ~60 amino acids, and the secondary structure of the protein forming a barrel structure, in which the walls are formed by beta sheets (in which hydrogen bonds are formed between two parts of the protein chain that can be far apart). These Ig domains were later found in many more proteins, creating an "Ig superfamily" that encompasses a multitude of immunologically important proteins [for example, T-cell (thymus-derived lymphocyte) receptors, major histocompatibility complex (MHC) molecules, and even adhesion molecules].

More detailed analyses of these monoclonal antibodies detected further differences in the variable regions of antibodies. There were stretches of amino acids that were more conserved between the different monoclonal immunoglobulins, and others that showed a particular high variability. The three conserved regions basically build the backbone of the variable region (the barrel) and were named framework regions (FR), while the three sections of amino acids that showed high variability formed amino acid loops with higher flexibility. They were called complementarity-determining regions (CDRs) because they participate in the actual binding of the antigen.

In summary, an antibody is built of two heterodimers, each consisting of one heavy and one light chain. Each of the two chains has a variable region, and together the two chains constitute the binding site for the antigen, which is formed by the complementarity-determining regions of the two chains. Thus, the antibody has two binding sites for antigen.

Variability. For a long time, the dogma "one gene–one protein" ruled molecular biology. This would mean two genes are required to encode one antibody, since an antibody consists of two different protein chains (heavy and light). The question then arises as to how one human being can make millions and millions of antibodies with different specificities that react against basically everything in this world when the human genome encodes only 30,000–40,000 genes. The solution came in 1979 when Susumu Tonegawa discovered that immunoglobulin genes are assembled from gene segments. Heavy chains are assembled from three different segment types, variable (VH), diversity (D), and joining (JH) segments (where H stands for heavy), to be rearranged into one functional VHDJH variable region. The rearrangement of the variable light-chain region is more simple; only two gene segment types are rearranged, VL and JL (where L stands for light), omitting the D region (**Fig. 2**).

In the human germline for the heavy chain (before rearrangement happens), there are about 50 VH genes, 27 D genes, and 5 JH genes, all lined up on the DNA and interrupted by stretches of noncoding DNA. All these gene segments then can be combined freely, as long as they join to make a VDJ region in the end. To get to this point, first a D region is attached to a JH region, making a DJH fragment. Next, a VH is rearranged to attach to a DJH fragment. The V, D, and J segments are flanked by small recognition sequences (called RSS for recombination signal sequences) that are mandatory for the rearrangement. These guarantee that the VH rearranges only to attach to the D, and the D to the JH, but not one JH to another JH, or a VH to a JH. On the molecular level, two recombination signal sequences are brought together and the intervening noncoding DNA is looped out. Two enzymes called RAG1 and RAG2 (recombination-activating genes) then cut the DNA at the outer ends of the coding sequences (VH, D, or JH). This cut is not precise and can add to the variability of an antibody. Before ligation (sewing them back together) of the two DNA ends, another enzyme (TdT: terminal deoxynucleotidyl transferase) can add random nucleotides. This too is a major contributor to add variability to the sequence and eventually the antibody. Together these nucleotides encode amino acids that had not been encoded by the germline DNA but increase the variability of antibodies extremely. While TdT and RAG1/2 are the only enzymes that are specific for B and T cells, there are many more enzymes involved in the process of rearrangement that are expressed ubiquitously. *See* DEOXYRIBONUCLEIC ACID (DNA).

The light chains are more simple in structure; they consist only of V and J segments, which are brought together in one simple rearrangement. In contrast to the heavy chain, there are two light chains to choose from, called the kappa and the lambda chain. In humans, about 50% of antibodies have a kappa light chain. The one heavy chain expressed can pair with only one type of light chain per cell, kappa or lambda. The order of the gene segments differs slightly between the kappa and lambda chain. While in both cases all V segments are together, the V segments in the kappa chain are followed by a cluster of five J segments and one constant region. The lambda locus has three available J segments with a compulsory lambda constant region in each.

The last way to increase variability is the ability of every heavy chain to combine with every single light chain when forming an antibody.

In summary, this makes a total of four ways to increase variability, which can be differentiated as combinatorial and junctional diversity. The free choice of picking V, D, and J segments for rearrangement and the fact that all generated heavy chains can pair with any light chain are considered combinatorial diversity. The junctional diversity is generated by a sloppy mechanism when cutting out the noncoding DNA, and by the addition of random nucleotides before ligation of the DNA. Four mechanisms, though, do not seem powerful enough to increase the variability to the amount that a few genes could cover millions of antigens. However, a running example with numbers will give a better picture. Assuming 51 VH, 27 D, and

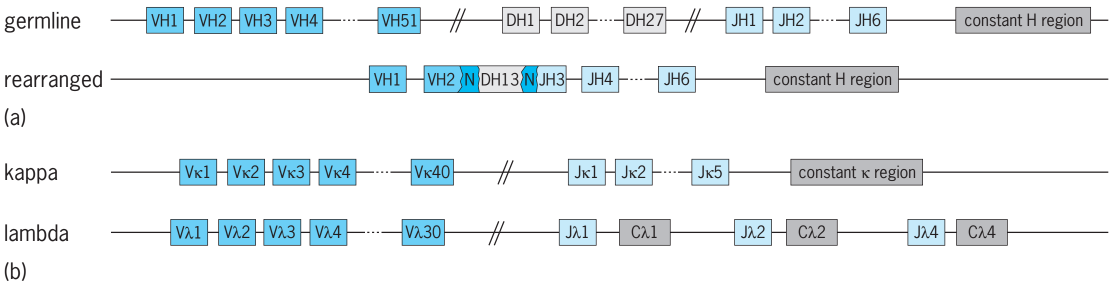


Fig. 2. Human immunoglobulin loci of heavy and light chains. (*a*) Heavy-chain immunoglobulin locus showing the lineup of the V, D, and J segments. The more complex constant region is depicted as a box (for more details, see Fig. 3). The germline configuration applies before rearrangement happens, or in a cell that does not undergo rearrangement (a liver cell, for example). In the example of a rearranged locus, VH2 has rearranged to DH13 and JH3. In the borders between the fragments, the cutting mechanism showed some sloppiness (wavy lines), and some N nucleotides have been added between the segments. Note that the intervening sequences from VH14 to DH12 and from DH14 to JH3 have been deleted from the genome. Double bars between segments indicate many kilobases of intervening noncoding sequence. (*b*) Light-chain immunoglobulin loci in germline configuration. While the kappa chain resembles that of the heavy-chain lineup (the only exception being that it does not have D elements), the lambda chain shows three constant regions with one linked J for each. Rearrangement to J always leads automatically to the expression of the linked constant region.

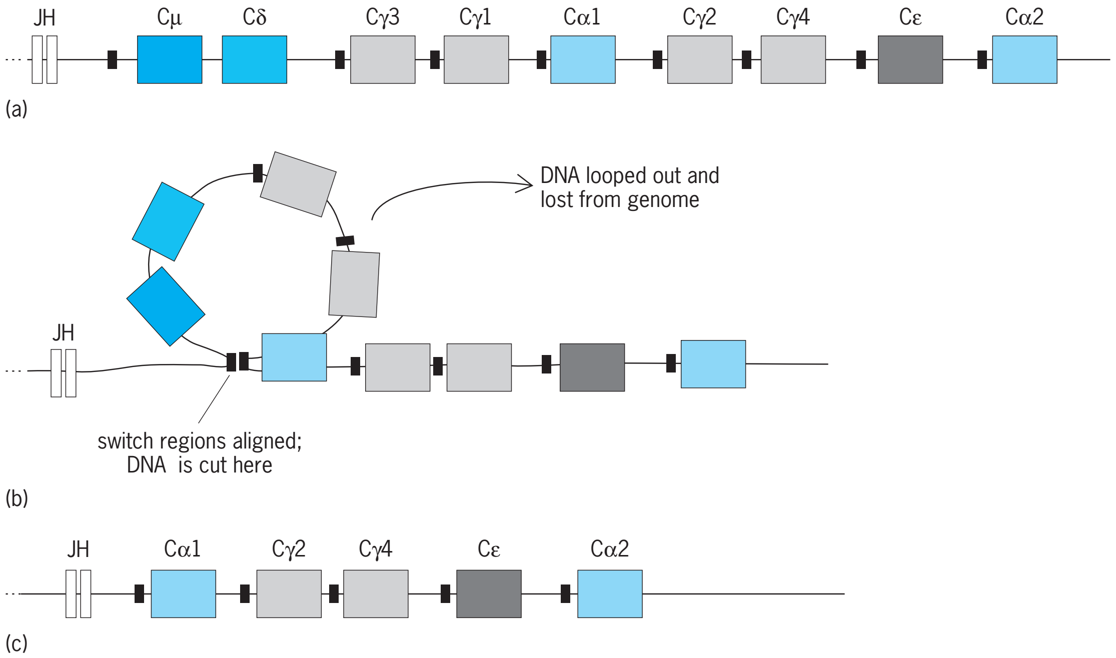


Fig. 3. Heavy-chain constant region. (*a*) Locus before class switch. The Greek letters in the constant-region names indicate the class of antibodies that will be expressed, μ for IgM, γ for IgG, etc. Some of the classes have multiple subclasses: in humans IgG actually is IgG1–4. Note that all classes have a switch region (little box) in front of their gene, except for Cδ, which is cotranslated with Cμ and then alternatively spliced. Thus it is possible to switch from IgM to any other class except IgD. (*b*) During class switch, the switch regions are brought into proximity, and DNA is looped out, cut, and religated. The intervening sequence is circularized and lost from the chromosome. (*c*) Constant region after class switch. Due to the loss of Cμ, the cell is never able again to make an IgM.

6 JH segments that can combine freely, the heavy-chain rearrangement alone gives $51 \times 27 \times 6 = 8262$ possibilities. If the light chain has $40\ \mathrm{V} \times 5\ \mathrm{J} = 200$ combinations, the free combination of 8262 heavy chains with 200 light chains gives $8262 \times 200 = 1.65$ million antibodies with different specificities. As we take into account the sloppiness of the process and the addition of N nucleotides, let us add an example: Assume that 10 different sequences are possible at each of the segment boundaries for any one combination of Vκ, Jκ, VH, D, and JH gene segments (this is an underestimate). This would multiply diversity by a thousandfold. Hence, the 1.65 million binding sites generated by combinational diversity alone is now amplified to 1.65 billion antibody specificities, starting out with less than 200 gene fragments.

Ig classes. While the light chain has only one constant region, the heavy chain has different constant regions available, representing different antibody classes (**Fig. 3**). These classes are called IgM, IgG, IgA, IgE, and IgD, and all have specialized functions. However, a B cell always starts out making IgM and IgD by default. After a cell has been activated and participates in an immune response, it can switch to other classes. This is a DNA recombination process that is very similar to VDJ rearrangement. Again, there are small regions that direct the reorganization of the DNA (here just called switch regions); there is one specific enzyme known to date that participates (AID: activation-induced cytidine deaminase); and as before, the DNA is looped, cut, and religated. Once IgG is made, the genetic information for IgM is lost and the cell can never again make IgM. Crucial in this class switch is that while the antibody changes the effector function by changing the class, it does not change the specificity.

Alternative splicing. Immunoglobulins not only are secreted as antibodies, but also serve a second purpose on the surface of B cells as B cell receptors (BCRs). The B cell receptor binds antigen, and once many of these are cross-linked by antigen, they signal the cell that it has been activated to participate in an immune response. But how can the same protein on one hand be secreted and thus be soluble in a watery solution (the serum), and on the other hand be anchored in the cell membrane and thus have a fatty region to stick to the membrane? The answer is alternative splicing, a process in which a long primary RNA is cut in different ways to encode for different proteins. In the case of immunoglobulins, alternative splicing determines if an immunoglobulin becomes a secreted antibody (the RNA is cut to encode for a hydrophilic protein tail so the antibody is soluble in serum) or a membrane-bound BCR (it encodes a stretch of lipophilic amino acids that can anchor the immunoglobulin in the membrane). This is the case for all classes, as all primary RNAs have the information for secreted and membrane-bound forms enclosed.

Another special case concerns the IgM and IgD classes. These two classes are actually encoded as a very long combined RNA transcript, and alternative splicing in this case not only determines whether it will be secreted but also whether IgM or IgD is made. All the other classes make only primary transcripts for their class. *See* CELLULAR IMMUNOLOGY; IMMUNOCHEMISTRY; IMMUNOLOGICAL ONTOGENY; RIBONUCLEIC ACID (RNA). Katja Aviszus

Bibliography. G. M. Edelman, Biochemistry and the sciences of recognition, *J. Biol. Chem.*, 279(9):7361–7369, 2004; T. J. Kindt, B. A. Osborne, and R. A. Goldsby, *Kuby Immunology*, 6th ed., W. H. Freeman, 2006; R. R. Porter, Structural studies of immunoglobulins, *Science*, 180:713–716, 1973; S. Tonegawa, Somatic generation of immune diversity: The Nobel lectures in immunology, *Scand. J. Immunol.*, 38:303–319, 1993.

Immunologic cytotoxicity

The mechanism by which the immune system destroys or damages foreign or abnormal cells. Immunologic cytotoxicity may lead to complete loss of viability of the target cells (cytolysis) or an inhibition of the ability of the cells to continue growing (cytostasis). Immunologic cytotoxicity can be manifested against a wide variety of target cells, including malignant cells, normal cells from individuals unrelated to the responding host, and normal cells of the host that are infected with viruses or other microorganisms. In addition, the immune system can cause direct cytotoxic effects on some microorganisms, including bacteria, parasites, and fungi.

Immunologic cytotoxicity is a principal mechanism by which the immune response copes with, and often eliminates, foreign materials or abnormal cells. Cytotoxic reactions are frequently observed as a major component of an immune response that develops following exposure to foreign cells or microorganisms. In addition, there is increasing evidence that cytotoxic reactions represent a major mechanism for natural immunity and resistance to such materials. In most instances, cytotoxicity by immune components involves the recognition of particular structures on the target cells; also, the targets need to be susceptible to attack by the immune components. Some cells are quite resistant to immunologic cytotoxicity, and this appears to represent a major mechanism by which they can escape control by the immune system.

There are a variety of mechanisms for immunologic cytotoxicity. The two main categories are antibody- and cell-mediated cytotoxicity. Within cell-mediated cytotoxicity, there is a multiplicity of effector cell types and mechanisms that can be involved.

Antibody-mediated cytotoxicity. Cytotoxicity by antibodies is usually complement-dependent. Certain antibodies are of the classes of immunoglobulins that have the ability to bind to the series of serum proteins that collectively are known as complement. Immunoglobulin M and some subclasses of immunoglobulin G (IgG1 and 3 strongly, IgG2 weakly, in humans) have the ability to bind complement components, and only these are cytotoxic. *See* COMPLEMENT; IMMUNOGLOBULIN.

As part of an immune response to foreign materials, cytotoxic antibodies are often induced. Moreover, most normal individuals have natural antibodies against a wide variety of cells and microorganisms, and some of these natural antibodies can be cytotoxic in the presence of complement. For cytotoxicity by antibodies to occur, the antibody needs to first bind to the relevant antigens on the target cells, and this is followed by binding of a series of complement components to the immune complex. After binding of the full array of complement components, a susceptible target cell is rapidly lysed. The mechanism for such damage is thought to be related to the creation by the complement components of pores in the surface membrane of the cells, with subsequent loss of intracellular contents. *See* ANTIBODY; ANTIGEN.

Cytotoxic T lymphocytes. Immune T cells are the best-known effectors of cell-mediated cytotoxicity. Upon immunization with cells that differ from the host in their major histocompatibility antigens and are recognized as foreign, cytotoxic T cells are generated which have a potent ability to lyse a variety of target cells bearing the foreign histocompatibility antigens. In addition, cytotoxic T cells can be generated against a variety of other foreign antigenic structures on the surface of cells. This can include antigens associated with tumors of the host, and also host cells infected with various microorganisms, particularly viruses. In general, cytotoxicity against cells with these other antigens is displayed only when the cells share the same major histocompatibility antigens with the responding host. Only some subpopulations of T lymphocytes have the ability to develop cell-mediated cytotoxicity, presumably because they contain the necessary intracellular machinery for such damage.

The actual mechanism of killing by immune T cells is not well defined, but it is clear that binding of the immune cells to the targets is a requisite first step, and that this is followed by a series of metabolic events both in the immune T lymphocytes and in the target cells that eventually leads to lysis of the targets. The immune T cells themselves are not damaged during the interaction, and have the ability to go on to kill other target cells. Such unidirectional cytotoxicity of only the target cells and not the effector cells is characteristic also of the other effector cells discussed below. *See* CELLULAR IMMUNOLOGY.

Macrophage-mediated cytotoxicity. Macrophages and the related monocytes, the principal scavenger cells in the body, have the ability not only to ingest foreign cells or materials but also to cause cytolytic or cytostatic effects on a variety of target cells. Cytotoxic macrophages appear to play a major role in the elimination of various microorganisms. They also have an ability to discriminate between tumor cells and normal cells and to selectively damage a wide variety of tumor cells. Macrophages and monocytes from normal individuals frequently have low levels of cytotoxic reactivity. They can be activated to a potently cytotoxic state, mainly by interaction with a product of immune T cells that is referred to as the macrophage-activating factor. In addition, interferon that is produced by immune T cells or by other cell types can activate macrophages to become cytotoxic.

As with cytotoxic T cells, the first step in the cytotoxic reaction by macrophages involves binding to structures recognized on the target cells. This appears to be followed by either of two types of metabolic events that can lead to damage of the targets. One major mechanism is the development in the activated macrophage of hydrogen peroxide and other reactive oxygen species, which can be very toxic against some microorganisms and other cells. Moreover, it appears that cytotoxic macrophages can release proteolytic enzymes which can cause damage and lysis of target cells. *See* BLOOD.

Natural killer cells. In the course of studies designed to find immune cytotoxic T cells in humans or animals with tumors, it was found that normal individuals also have lymphocytes with the ability to recognize and kill a variety of tumor cells. Such effector cells have become known as natural killer (NK) cells. It has been shown that natural killer cells represent a discrete subpopulation of lymphocytes which vary morphologically from typical lymphocytes and are known as large granular lymphocytes, because of their larger size and the presence of characteristic granules in their cytoplasm. Natural killer cells have the ability to recognize and lyse a wide variety of target cells, including tumor cells, some cells of the embryo and of normal bone marrow and thymus, and a variety of microorganisms. Normal fibroblasts or other cells, upon infection with viruses, become highly susceptible to lysis by natural killer cells. There is increasing evidence that natural killer cells play a major role in natural resistance against tumors, particularly against the metastatic spread of tumor cells. They have also been shown to be involved in natural resistance against infections by certain viruses, fungi, and parasites. In addition, they appear to play an important role in the rejection of grafts of foreign bone marrow cells. *See* TRANSPLANTATION BIOLOGY.

Granulocytes. Granulocytes are the most common type of white blood cell in the blood and tissues. These cells have the ability to mediate immunologic cytotoxicity against tumor cells and microorganisms. During the course of an immune response, proteins produced by immune T cells or other cells can activate the granulocytes to become highly cytotoxic. In addition, some components of microorganisms themselves have the ability to activate granulocyte-mediated cytotoxicity. A major mechanism for such cytotoxicity by granulocytes is the release of hydrogen peroxide or other reactive oxygen species, similar to that described above for macrophages.

Antibody-dependent cell-mediated cytotoxicity. Immunologic cytotoxicity also can be mediated by a collaboration between immune or natural antibodies and certain immunologic effector cells. It is known as antibody-dependent cell-mediated cytotoxicity, and it appears to be a potent mechanism for elimination of some microorganisms; it may also be involved in the destruction of tumors and other foreign cells. The initial event in antibody-dependent cell-mediated cytotoxicity is an attachment of antibodies, usually immunoglobulins of the G class, to target cells. In turn, certain types of effector cells are bound by means of cell surface receptors for a portion of immunoglobulin G that is shared by most antibodies (the Fc moiety). After binding of the effector cells to the targets, a series of metabolic events occurs which leads to the lysis of the targets. Antibody-dependent cell-mediated cytotoxicity can be mediated by a variety of lymphoid cells, particularly macrophages, natural killer cells, and granulocytes. Therefore, this represents an alternative mechanism by which these cytotoxic effector cells can recognize and destroy target cells. *See* IMMUNOLOGICAL DEFICIENCY; IMMUNOLOGY.

Ronald B. Herberman

Bibliography. E. S. Golub (ed.), *The Cellular Basis of the Immune Response: An Approach to Immunobiology*, 1980; R. B. Herberman (ed.), *NK Cells and Other Natural Effector Cells*, 1982; D. P. Stites and A. I. Terr, *Medical Immunology*, 9th ed., 1997.

Immunological deficiency

A state wherein the immune mechanisms are inadequate in their ability to perform their normal function, that is, the elimination of foreign materials (usually infectious agents such as bacteria, viruses, and fungi). Immune mechanisms are also responsible for the rejection of transplanted organs. The processes described above are accomplished by white blood cells known as lymphocytes, of which there are two major types, T lymphocytes (thymus-derived) and B lymphocytes (bone marrow–derived). *See* CELLULAR IMMUNOLOGY.

Development of competent immunity. T and B lymphocytes are derived from immature cells known as stem cells, which normally reside in the bone marrow. In order to perform their protective functions effectively, both T and B lymphocytes must undergo processes of maturation. For T cells, this development takes place in the thymus gland, while B cells mature in the bone marrow. *See* IMMUNOLOGICAL ONTOGENY.

In general, a given lymphocyte reacts to only one antigen. From this specialized lymphocyte descends a family of cells, which respond to that specific antigen. It has been estimated that normal humans have the capacity to recognize and respond to over a million different antigens; thus, a million different lymphocyte families or clones could be generated. Human lymphocytes recognize different antigens by having specific receptors on their cell surfaces. A lymphocyte responds to the antigen only if there is a fit between the lymphocyte receptor and the shape of the antigen. The reaction is similar to the fit of a lock and key. *See* ANTIGEN.

Upon encountering an antigen, the responsive lymphocytes increase markedly in number; thereafter, the two types of lymphocytes respond differently. The T-cell family forms a class of helper cells which enhance the response to the antigen, and another set, called suppressor-killer cells, which perform a double duty. The killer function gives a cell the ability to destroy target cells, whereas the suppressor capacity slows down the immune response when it is sensed that the emergency is over and the immune system can be put to rest. The response to antigen by the B lymphocytes is to generate a large number of cells that manufacture and secrete proteins (immunoglobulins) that can physically combine with the antigen. *See* IMMUNOLOGIC CYTOTOXICITY.

The five major classes of immunoglobulins, IgM, IgG, IgA, IgE, and IgD, are known as isotypes. IgM is thought to be the first line of defense and responds very quickly to infections. It is mainly contained in the bloodstream, where it can readily neutralize agents attempting to gain entrance through the blood. IgG is the immunoglobulin which confers long-lasting protection and is transferred to the fetus through the placenta. In this way, the baby is protected by the mother's IgG for the first few months of life. Other Ig classes do not cross the placenta.

IgA is localized primarily in the lungs and intestine, where it protects from organisms which are inhaled or eaten. IgE causes the symptoms of hay fever and asthma. It is considered to have evolved as a component of the protective mechanism against parasites. For instance, its release causes the intestine to contract violently, which may expel an intestinal worm. *See* IMMUNOGLOBULIN.

When either a killer T cell or an immunoglobulin molecule combines with the antigen, other body reactions are brought into play to make the elimination of the antigen more effective and complete. T cells secrete a number of molecules known as lymphokines which attract other cells that are also able to destroy target cells and in general make the body more resistant to the antigenic material. The immunoglobulins can combine with another set of proteins, known as the complement system. The complement cascade consists of 11 proteins which act in a sequential manner to destroy whatever target is coated by the immunoglobulin. *See* COMPLEMENT.

Causes. Immunodeficiency diseases (IDDs) result from a failure at any point in the complex set of interactions described above. In general, the diseases are due to absence of cell populations; failure of cells to mature; failure to secrete the products (such as lymphokines) necessary for effective cell interactions; or failure of accessory cell populations or protein systems (for example, complement) which are necessary for the complete competence of B-cell immune function. An immune response begins when immune system cells are triggered by an antigen to which the individual has been immunized. This response consists of recognition of the stimulus (often an infection) by external cell receptors which leads to a series of internal cell signals that initiate the process of cell synthesis and secretion. These responses to external stimuli are known as signal transduction. Defects of enzymes vital in signal transduction account for many immunodeficiency diseases. Some of the diseases are carried on the X chromosome and affect only males, being carried by females. *See* SEX-LINKED INHERITANCE.

In recent years, impressive progress has been made in defining causes of IDDs. The specific protein defect and the chromosomal location of its gene have been defined (**Table 1**).

Symptoms. It is apparent from a consideration of the functions of the immune system that the prime symptom of immunodeficiency would be an increased susceptibility to infections. Since T cells and B cells differ in the effectiveness with which they protect against various kinds of organisms, some insight into which system (T cell or B cell) is primarily

TABLE 1. Genes involved in immunodeficiency disease

Disease	Gene product affected	Chromosome location of gene
Severe combined immunodeficiency (X-linked)	γ_c; component of several cytokine receptors	Xq13.1–13.3
Severe combined immunodeficiency (non-X-linked)	Jak3, necessary in signal transduction	19p13.1
Ataxia-telangiectasia	Ataxia-telangiectasia-mutated (ATM), which is part of mechnisms that repair damaged DNA, a process necessary for the continued production of normal cells	11q23.1
Wiskott-Aldrich syndrome	Wiskott-Aldrich syndrome protein (WASP), part of a complex protein system which is critical for antigen recognition and signal transduction	Xp11.23
Adenosine deaminase deficiency	Adenosine deaminase, necessary for catabolism of cellular breakdown product; in its absence, toxic metabolites develop which destroy and B cells	20q13-ter
Nucleoside phosphorylase deficiency	Nucleoside phosphorylase, an enzyme with functions similar by to adenosine deaminase but acts at a different site in the cellular breakdown metabolic pathway. the accumulated toxic metabolites destroy T cells primarily	14q13.1
DiGeorge anomaly*	Specific protein defect not defined as yet	22q11.2[2]
X-linked hypogammaglobulinemia	Bruton tyrosine kinase, an intracellular protein important in cell signaling pathways	Xq21.3–22
X-linked hyper IgM syndrome	CD40 ligand (CD40L), a protein on T cells which reacts with the B-cell membrane protein CD40, causing B cells to switch from IgM to IgG and IgA synthesis	Xq26.3–27.1

*Some cases are associated with deletions in chromosome 10p.

TABLE 2. Combined T- and B-cell diseases

Disease	Distinguishing clinical features*	Remarks
Severe combined immunodeficiency	Total lack of hair or severe skin infections may be an early sign	Often fatal by age 2
Short linked dwarfism or cartilage hair hypoplasia	Very short legs Characteristic x-rays Children have fine thin hair Double-jointed	There may be susceptibility only for chickenpox
Ataxia-telangiectasia	Prominent blood vessels in white of eyes, ear lobes, and cheeks Poor coordination Ultimately mental retardation Poor tolerance to x-ray therapy	High incidence of cancer
Wiskott-Aldrich syndrome	Severe eczema Frequent severe herpes infection Low blood platelets cause excessive bleeding X-linked	High incidence of cancer
Adenosine deaminase deficiency	May have characteristic x-rays with disease involving ribs, pelvis, arms, and legs	B-cell deficiency may be present early or occur with time; first disease to be treated by gene therapy

*For general symptoms common to all diseases, see text.

affected can be gained by considering the kind of organisms involved. T cells primarily fight off infections caused by yeasts, viruses (causing chickenpox or measles), fungi, and certain bacteria, such as those causing tuberculosis. B cells are particularly effective against most bacteria (for example, staphylococci, streptococci, and pneumococci). There is also evidence that tumor cells are eliminated by cells of the immune system (probably T cells and a related type known as the natural killer cell). In any event, the incidence of some types of cancer (largely of the reticuloendothelial system) in immune-deficient populations is 10,000 times that of normals.

Many of the organisms to which people are constantly exposed do not ordinarily have the capability to cause infections in immunocompetent individuals because these organisms are so weak that they cannot establish themselves in normals. In immunodeficients, however, they can cause fatal infections. One such organism is a protozoon, *Pneumocystis carinii*, which can cause debilitating pneumonia. In other situations, agents which usually cause only mild infections lead to fatal disease. Chickenpox, cytomegalovirus, and measles will cause fatal pneumonia in deficient individuals. Chronic parasitic infections leading to severe diarrhea are also characteristic of this deficient group. Thus, immunodeficient individuals respond severely or fatally to very mild or ordinarily innocent infectious agents.

Specific diseases. There are many varieties of immunological deficiency diseases. In general, it is sufficient to think of the diseases as involving primarily the T-cell system, the B-cell system, or both cell lines. Involvement of both T and B cells is known as combined immunodeficiency disease and is the most severe form. It is often fatal, usually before the age of 2 years. Defects which involve the T cells but not the B cells are the next most severe. Pure B-cell deficiencies are the mildest and are known as hypogammaglobulinemias. In some of the variants, abnormalities of the red blood cells, white blood cells, or blood platelets are associated. Occasionally, one also finds biochemical abnormalities and disorders of the bones (**Tables 2–4**).

It is possible to have single immunoglobulin deficiencies. The most common immune defect is an isolated deficiency of IgA, thought to occur in approximately 1 out of 600 people. In most individuals with selective IgA deficiency, compensatory mechanisms keep those affected individuals from having serious disease. In fact, many people with IgA deficiency are unaware of their abnormality.

Most of these problems are due to inherited faults. In general, immunodeficiency states are recessive traits, inherited as either autosomal or X-linked traits. As a result, approximately 25% of all children in such families can be expected to be affected. In the case of X-linked disorders, females will not be affected but

TABLE 3. Diseases predominantly involving T cells*

Disease	Distinguishing clinical features*	Remarks
Nucleoside phosphorylase deficiency	Severe anemia is often associated	
Chronic mucocutaneous candidiasis	Skin lesions involve areas on face, and trunk, nails, and often lead to severe disfigurement	Often associated with diseases of endocrine glands causing hypothyroidism and diabetes May be due to biotin defect
DiGeorge anomaly	Associated with characteristic facial appearance (receding chin, low-set ears, widely spaced eyes), severe heart disease, and low calcium	Occurs with all degrees of severity and often has associated B-cell defect

*For general symptoms common to all diseases, see text.

TABLE 4. Diseases predominantly involving B cells

Disease	Distinguishing clinical features	Remarks
X-linked hypogammaglobulinemia Congenital hypogammaglobulinemia (Bruton's syndrome)	Disease present at birth Tonsils and lymph nodes very small or absent	B cells absent Occasionally associated autoimmune phenomena (rheumatoid arthritis, skin and muscle disorders) Responds well to gamma globulin injections or infusions
Common variable immunodeficiency	Disease usually begins later in life	Variable degree of T-cell defect more common B cells present but nonfunctional One variety associated with thymus gland tumor (thymoma)
X-linked hyper IgM	Neutropenia common IgG and IgA absent, IgM normal or at elevated levels	Pneumocystis pneumonia may occur Low platelet number: severe intestinal and liver disease may be seen
Chronic lymphatic leukemia	Large spleen	Cancer of the lymphocytes; 25% have B-cell deficiency
Multiple myeloma	Gammaglobulin levels paradoxically very high Disease commonly produces holes in bones due to involvement by cancer	Deficiency due to replacement of effective immunoglobulin molecules by the abnormal myeloma protein

will carry the trait. In this instance, the 25% figure is derived from a 50% incidence in male births, which occur 50% of the time.

Immunodeficiency can also be acquired as a complication of other disease processes. One of the most common forms of deficiency is caused by aggressive treatment of leukemia. Unfortunately, when the cancer cells are destroyed, the normal lymphoid cells are killed as well. As yet there are no known agents which are selective only for cancer cells; in most cases they preferentially react against the malignant cell, but in the high dose required, the normal cells ultimately are lost, too. Another cause of induced immunodeficiency is seen with transplant rejection therapy. To prevent organ rejection, drugs which destroy lymphocytes must be administered. The strategy is to employ enough medicine to prevent rejection without completely losing all lymphocyte function. This fine line is frequently crossed, however; and serious, even fatal, infection can complicate organ transplantation. *See* LEUKEMIA; TRANSPLANTATION BIOLOGY.

A fascinating variety of immunodeficiency is secondary to virus infections. Certain viruses, such as the Epstein-Barr virus (EBV), which causes infectious mononucleosis, infect lymphocytes. Involvement of the lymphoid system is nearly always only temporary, and most individuals can clear the virus after a few weeks. In a small number of individuals, probably as a result of a genetic defect, the virus cannot be eliminated, and a chronic infection of B cells leads to the loss of normal lymphocyte function. Hypogammaglobulinemia has been described after EBV infection. An overwhelming immunodeficiency disorder, known as the acquired immune deficiency syndrome (AIDS), has been observed primarily in homosexuals, intravenous drug users, and hemophiliacs. The cause is infection with a virus termed the human immunodeficiency virus (HIV). *See* ACQUIRED IMMUNE DEFICIENCY SYNDROME (AIDS); EPSTEIN-BARR VIRUS.

Immunodeficiency can be observed secondary to dietary deficiency. Two main varieties are seen. In protein-calorie malnutrition, serious deficiency primarily involving the T-cell system predisposes affected individuals to overwhelming infection by the agents of measles or tuberculosis. In the second variety, deficiency of single substances is the cause; the two most commonly observed deficiencies are those of zinc and biotin. It should be pointed out that deficiency disorders are caused by extremely severe dietary restriction or rare absorption defects, and that among the general population it is not necessary to enhance the diet with massive dietary supplements to enhance immunity. In fact, ingestion of tremendous amounts of many supposedly innocuous foods or vitamins can be dangerous. *See* BIOTIN; MALNUTRITION; VITAMIN.

Tests of immune function. The capability of the immune system can be precisely estimated by examination of the blood. The T cells and B cells can be stained with special reagents that are highly specific for each cell type. It is even possible to differentiate between the special varieties of T cells, that is, the helper and the suppressor-killer types. Helper cells are designated CD4, and suppressor-killers are known as CD8 according to their reactivity with corresponding monoclonal antibodies. B cells are designated CD19. The different classes of immunoglobulins are measured by chemical tests. By taking samples from thousands of normal persons, values in healthy individuals have been established. One can then compare the values observed in a patient to determine if a deficiency exists. *See* MONOCLONAL ANTIBODIES.

Sometimes the cells or immunoglobulin proteins are present in normal amounts, but their functional capability is weakened. That is, they will not combine with, or respond to, antigens. Special tests can be employed to actually mimic an immune response in the test tube. These tests make it possible to determine whether a person's lymphocytes can behave in a normal manner when stimulated by antigens.

A form of immunodeficiency associated with enzyme defects has been described. The enzymes involved are adenosine deaminase and nucleoside phosphorylase (NP), necessary, respectively, for the catabolism of the purines adenosine and inosine. In the absence of these enzymes, abnormal metabolites accumulate and are toxic to the lymphocytes. The T cells are more susceptible to the toxicity, and hence the defect primarily involves that cell line. Ultimately, in adenosine deaminase deficiency, B cells as well become involved, so that combined deficiency occurs with time. For unknown reasons, the B cells remain uninvolved in nucleoside phosphorylase deficiency. These biochemical defects are inherited as autosomal recessive traits.

Further information can be obtained by studying the sites of the immune reaction by taking a sample of a lymph node. A small sample of the thymus gland can also be studied. Lymphoid tissues which are functioning normally show characteristic features under the microscope.

Diagnosis of immunodeficiency diseases can be made in utero in many cases. This has been greatly facilitated by molecular definition of the defect and determination of the chromosomal localization of the gene involved. In addition, fetal blood samples can be obtained and the cells tested in the same manner as cells taken from patients in the clinic.

For X-linked diseases, females at risk (aunts and sisters of patients) can be studied and the carrier status determined.

Treatment. The treatment depends upon the defect. Pure B-cell disorders are the simplest to treat and usually respond to immunoglobulin injections or infusions. These are required for the rest of the life of the individual.

Pure T-cell disorders are very rare but can be treated by thymus transplantation. It is not necessary to place the organ in its natural site (in the chest underneath the breastbone); the organ can be transplanted in a large muscle. The rich blood supply of the muscle will bring the stem cells from the bone marrow to the transplant, where maturation of T cells occurs, allowing replacement of the missing cell population. Since individuals with T-cell deficiency are unable to reject organs, it is easy to perform the transplant without any concern about rejection of the graft. *See* THYMUS GLAND.

In combined immune deficiency states, it is usually necessary to provide both B and T cells. This is best accomplished by a bone marrow transplant. The theory is that the bone marrow provides stem cells which can mature in the recipient. However, it is now apparent that the bone marrow contains mature as well as immature cells, and exactly how much of the reconstitution is due to immature cells and how much to mature cells is presently unknown. It is not possible to transplant bone marrow between any two individuals chosen randomly. The reason is that the mature T cells in the bone marrow suspension can react vigorously against any foreign substance. In the case of a transplant, that foreign substance would be the recipient and the donor bone marrow could actually “reject” the individual.

This is exactly the same reaction which occurs when an immunocompetent individual rejects a transplant. It is called host-versus-graft reaction; in this case, the graft dies. In the bone marrow situation, the marrow is the graft, and when it rejects the recipient, it is called graft-versus-host reaction. This complication can be fatal for the host. Graft-versus-host disease is controlled in two ways. First, the donor and recipient must be matched as closely as possible according to bone marrow types. In principle, the matching is similar to matching people for blood transfusions but, in practice, is much more difficult because the lymphocyte system is so much more complex than the red blood cell system. There are only four major types of blood, and it is relatively easy to find compatible types among the population at large. There are thousands of lymphocyte variants, and in general, only family members can be considered as possible donors. As a practical matter, usually only brothers or sisters are checked, but even then, a match will be found only 25% of the time. On a rare occasion, a parent or more distant relative may match, especially if there has been marriage among close relatives (for example, cousins). The chance of finding a match between two unrelated individuals in the population at large is approximately 1 in 10,000. *See* BLOOD GROUPS.

It is also possible to eliminate the mature T cells from the marrow before transfusing it into the patient. This strategy has been used successfully in a number of transplants between parent and child. Because of the nature of the transplantation antigens, in nearly all cases, a parent will be matched to the child for one-half of all antigens. (Each person carries two sets of transplantation antigens—one set is inherited from each parent.) If a bone marrow transplant can be successfully accomplished, the individual can be expected to be cured.

Umbilical cord blood from closely matched donors is now used with great success. The T cells from umbilical cord blood are much less aggressive in causing graft versus host disease.

With the cloning of genes which cause immunodeficiency disorders, genetic reconstitution becomes possible. Adenosine deaminase (ADA) deficiency was chosen for the first clinical attempt. The gene was well suited for this trial, since ADA is a single-chain enzyme and its production does not require strict control. Thus, complicated problems of protein assembly are avoided, and increased levels are not harmful to the patient. Stem cells were isolated from the umbilical cord blood of three patients with ADA deficiency, the gene inserted, and the stem cells infused back into the patients. After 2 years, genetically corrected cells were still present in the patients’ blood and showed levels of ADA cDNA at 1/1000 to 1/10,000. However, injectable ADA which had been given to support the patients until sufficient enzyme could be produced by the corrected cells could not be decreased, so the levels of enzyme produced are not sufficient to correct the defect, as assessed 4 years after the cell infusion.

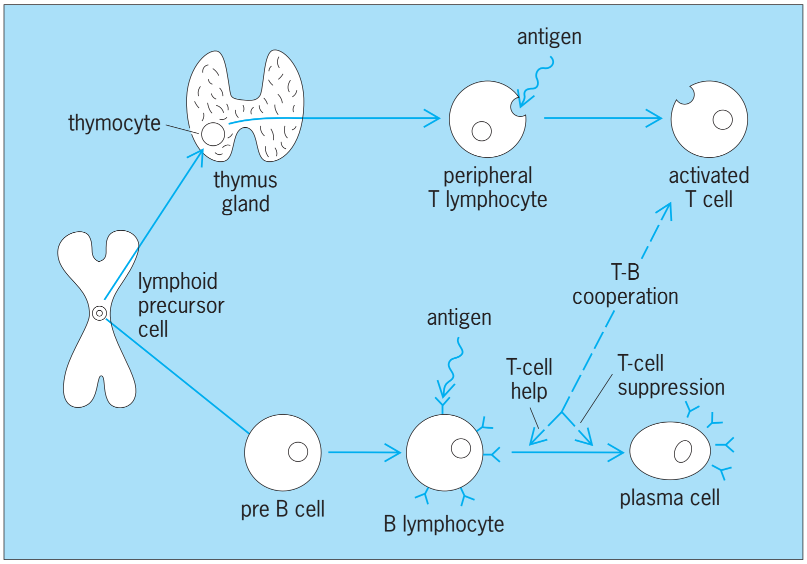
It was stated above that lymphocytes secrete various molecules which augment the immune response. Some of the substances have been purified and can be made in large quantity either by artificial synthesis or by gene-cloning methods. Substances which have been used include interleukin-2 (IL-2), interferon, and many of the hormones elaborated by the thymus gland (such as thymosin, thymopoietin, and TP-5). *See* GENETIC ENGINEERING; IMMUNOLOGY; THYMOSIN. Richard Hong

Bibliography. A. Fischer and B. Malissen, Natural and engineered disorders of lymphocyte development, *Science*, 280:237–242, 1998; F. S. Rosen et al., Primary immunodeficiency diseases, *Clin. Exper. Immunol.*, 109 (suppl. 1):1–28, 1997; E. R. Stiehm (ed.), *Immunologic Disorders in Infants and Children*, 4th ed., W. B. Saunders, Philadelphia, 1996; K. I. Weinberg (ed.), Molecular basis of lymphoid immunodeficiency, *Sem. Hematol.*, 35(4):280–367, 1998.

Immunological ontogeny

The origin and development (ontogeny) of the lymphocyte system, from its earliest stages to the two major populations of mature lymphocytes: the thymus-dependent or T lymphocytes, and the thymus-independent or B lymphocytes. The T lymphocytes carry out those aspects of function which are called cell-mediated immunity, including graft rejection, elimination of tumor cells, and delayed hypersensitivity. B cells are responsible for humoral or antibody-mediated immunity. *See* IMMUNITY.

For both systems, development or differentiation proceeds in discrete stages (see **illus.**). In the first stage, pluripotent hematopoietic stem cells, which originate in the yolk sac in the embryo and then successively in fetal liver and fetal bone marrow, develop into precursor cells committed to becoming T or B cells. Hematopoiesis in human fetal liver begins at about 4 weeks of gestational age and in fetal bone marrow after 20 weeks. *See* HEMATOPOIESIS.



Schematic representation of the differentiation of the immune system.

Ontogeny of T lymphocytes. The thymus plays a strategic role in the development of T lymphocytes. It arises from the third and fourth pharyngeal pouches, and between the seventh and tenth week of gestation it descends from the neck into the anterior mediastinum. The thymus is highly organized with a network of reticuloepithelial cells. T-lymphocyte precursor cells are attracted into the thymus where, under the influence of this microenvironment, they undergo rapid proliferation and maturation. In humans, stem cells enter the thymus beginning in the eighth week of gestation and then begin to undergo differentiation. The epithelial cells within the thymic stroma are primarily responsible for this development, because they contain hormonelike substances which trigger T-cell differentiation. By the ninth or tenth week, fetal thymocytes proliferate in response to phytohemagglutinum and allogeneic cells, indicating T-cell maturation. These cells also begin to express a variety of cell-surface markers, which parallels developing immunocompetence. From the thymus, these maturing T cells are exported to the peripheral lymphoid tissues, including lymph nodes, gut, spleen, and peripheral blood. In the developing peripheral lymphoid tissues, these T cells populate the deep and midcortical areas of the lymph nodes and periarteriolar regions of the spleen. Immunocompetent T cells can be identified in these tissues by the time of 13–14 weeks' gestation. *See* THYMUS GLAND.

As maturation proceeds, the T cells become specialized and different subsets can be identified through the detection of different cell-surface antigens by monoclonal antibodies or by measuring distinct functions. These functions include T-helper-cell activity which, on interacting with B cells, promotes their maturation into antibody-secreting cells, and T-suppressor-cell activity, which regulates this response.

Ontogeny of B lymphocytes. The developmental pathway of B lymphocytes has been studied in greatest detail in the chicken. In this species, the maturation of B cells begins with migration of progenitor cells to a specific microenvironment within the bursa of Fabricius (an outpouching of the hindgut, hence the term B or bursa-dependent lymphocytes). In mammals, there does not appear to be a discrete and specialized environment analogous to the bursa of Fabricius. The bone marrow seems to be the major site of B-cell differentiation in mature mammals. Similarly to the T cells, B lymphocytes pass through several differentiation steps in their ontogeny, which may be characterized by the expression of specific markers.

The earliest B cells identified in fetal liver are pre-B cells. As the cells mature, they express immunoglobulin M (IgM) and subsequently IgD on their surface. At this stage, the cell is a specific, competent B cell ready to interact with an antigen. In the course of B-lymphocyte differentiation, diversity of immunoglobulin classes is generated by an orderly switch from IgM to IgG to IgA with expression of the respective immunoglobulin on the cell surface.

This ontogenic sequence is supported by the expression of constant-region heavy-chain genes. *See* IMMUNOGLOBULIN.

Cells bearing surface IgM, IgG, and IgA have been demonstrated by immunofluorescence in peripheral blood, liver, spleen, and bone marrow at 11–12 weeks' gestation, and cells containing cytoplasmic immunoglobulins by 20 weeks in human fetuses. Synthesis of IgM has been noted as early as 10–11 weeks' gestation, and synthesis of IgE and IgG shortly thereafter. *See* IMMUNOFLUORESCENCE.

For B cells, the surface immunoglobulin serves as a specific receptor for antigen. Once activated by antigen and following interaction with T lymphocytes and monocytes, B lymphocytes terminally differentiate into mature antibody-secreting plasma cells. The immunoglobulin-producing cells differentiate relatively early in gestation, and the human fetus can produce significant amounts of antibody if stimulated by antigen. The development of a fetal antibody response to a specific antigen depends on the nature of the antigen and on the maturation of specific clones of antibody-producing cells, or may be limited by the ability of fetal monocytes to process certain antigens. Furthermore, active infection of the fetus may hasten or retard the process of developing immunocompetence.

Immunoglobulin gene expression. The initial step in B-cell differentiation is a genomic rearrangement of the immunoglobulin gene complex that brings a variable-region gene segment into the site of a constant-region gene segment. Once the immunoglobulin genes have undergone the rearrangements that are necessary to encode a complete molecule, immunoglobulin is synthesized. On the surface of the cell, the immunoglobulin can bind antigen and signal the next step in differentiation. This involves a switch from surface to secreted immunoglobulin. *See* IMMUNOGENETICS.

Diversity of immunoglobulin classes with different heavy chains (IgM, IgG, IgA, IgD, and IgE) is generated through a heavy chain constant region switching mechanism. Analysis of the heavy-chain locus of both humans and mice has demonstrated that the eight different types of heavy-chain constant-region genes are encoded on a single stretch of DNA. The signals responsible for heavy-chain class switch are not understood. The different classes of heavy chain confer different biological properties on the immunoglobulin molecule. *See* CELLULAR IMMUNOLOGY.

Erwin W. Gelfand

Bibliography. A. Baumgarten and F. Richards (eds.), *CRC Handbook Series in Clinical Laboratory Science*, vol. 1, 1978; M. D. Cooper et al., The functions of the thymus system and the bursa system in the chicken, *J. Exp. Med.*, 123:75–102, 1966; E. W. Gelfand and H.-M. Dosch (eds.), *Biological Basis of Immunodeficiency*, 1980; A. R. Lawton et al., Ontogeny of B-lymphocytes in the human fetus, *Clin. Immunol. Immunopath.*, 1:84–93, 1972; P. Leder, Genetic control of immunoglobulin production, *Hosp. Pract.*, 18:73–82, 1983; O. Stutman, Two main features of T-cell development: Thymus traffic and post-thymic maturation, *Contemp. Top. Immunobiol.*, 7:1–46, 1977.

Immunological phylogeny

The study of immunology and the immune system in evolution. All vertebrates can recognize and respond to nonself-molecular configurations on microorganisms, cells, or organic molecules by utilizing a complex recognition system termed the immune response. The presence of lymphocytes and circulating antibodies has been documented in all extant vertebrate species, ranging from the ostracoderm-derived cyclostomes, such as hagfish and lampreys, to mammals. However, the existence of induced, specific reactions directly homologous to the immune repertoire of vertebrates has not been clearly established in invertebrates. Studies have been done which focused on the presence and characterization of cells of the T and B lymphoid series in lower vertebrate species; the nature of the antibodies produced by lower vertebrate species and their relationship to well-characterized immunoglobulin classes of mammals; the extent of diversity in the antibodies of lower vertebrates with particular respect to the degree of variable-region heterogeneity within lower species compared with that exhibited by mammals; and the deoxyribonucleic acid (DNA) sequence of genes encoding variable regions of immunoglobin of lower species. Studies designed to pinpoint the origin of immune diversification have been actively pursued in the protochordates, such as tunicates, and there is also considerable interest in the defense mechanisms of invertebrate phyla not directly related to the evolutionary stream which gave rise to the chordates.

Cellular immune responses. The role of phagocytic cells in engulfing foreign pathogens has been documented in virtually all metazoan organisms. Phagocytic cells possess a limited capacity to discriminate self from nonself, and this is due in part to the presence of lectins (molecules capable of binding specifically to various sugars) on their surface. This conclusion has been directly shown for phagocytic cells of the oyster, and apparently correlates with the fact that most invertebrates possess lectins of various specificities circulating in their hemolymph. Although there is no evidence to suggest that invertebrate lectins and vertebrate immunoglobulins are homologous structures, sufficient diversity exists within lectins of certain species to indicate that these types of molecules and their cellular expression on phagocytes might serve as a primitive and universal recognition mechanism. *See* IMMUNOGLOBULIN; LECTINS.

Colonial tunicates (protochordates) have shown the capacity to reject allografts, a phenomenon comparable to T-cell–mediated graft rejection of true vertebrates, although the existence of true lymphocytes in protochordates has not been firmly established. All true vertebrates possess cells clearly recognizable as lymphocytes and can carry out T-cell functions, such as graft rejection, and show the capacity of B cells to synthesize and secrete immunoglobulins. The teleost fish and amphibians possess lymphocytes of distinct functional and cell-surface–maker phenotypes which resemble T and B lymphocytes of birds and mammals. In addition, functional assays indicate that teleost fishes and amphibians possess helper T lymphocytes which are capable of interacting specifically with B cells in the production of circulating antibody. Lampreys and hagfish lack discrete lymphoid organs characteristic of higher vertebrates but contain diffuse collections of lymphoid cells in the area of the gills, which may correspond to a primitive thymus, and in the gut, which may correspond to the spleen of higher vertebrates. By contrast, all placoderm-derived vertebrates (sharks, bony fishes, amphibians, reptiles, birds, and mammals) have clearly demarcated thymuses and spleens. True lymph nodes are not present in vertebrate species more primitive than mammals, but birds possess aggregates of lymphoid tissue probably serving a similar function. The thymus is a primary lymphoid organ which generates cells of the T-lymphocyte series. In birds, a discrete organ, the bursa of Fabricius, is the primary source of B cells. The precise anatomical location of the bursa is not known in mammals or vertebrate species distinct from birds, but some sort of functional equivalent must exist. *See* LYMPHATIC SYSTEM; THYMUS GLAND.

B cells of teleosts and amphibians express surface immunoglobulin, which has been isolated and characterized. It resembles in overall properties the 7S IgM (IgMm) which is associated with the membrane of B lymphocytes of mammals. Mammalian B lymphocytes express, in addition, IgD membrane immunoglobulin. This has not yet been conclusively shown on the surfaces of B lymphocytes of teleost and amphibian species. The nature of the antigen receptor on T lymphocytes is not exactly known, but studies with T lymphocytes of teleost fish suggest that a molecule related to immunoglobulin might serve as the specific antigen receptor on these cells. *See* CELLULAR IMMUNOLOGY.

Immunoglobulins of lower species. Humans possess five major classes or isotypes of immunoglobulin: IgG, IgM, IgA, IgE, and IgD. These classes are defined by the type of heavy chain (γ, μ, α, ϵ, or δ) which occurs in disulfide-bonded association with light chains (either of κ or λ type). The usual structure of an immunoglobulin consists of two heavy chains joined to two light chains; in order to form the combining site to antigen, a light chain must be associated with the heavy chain. The combining site for antigen is made up by interaction of the variable (V) regions of a light chain in association with that of a heavy chain. A molecule consisting of two light (L) and two heavy chains would thus have two combining sites. The variable regions constitute approximately the first 100 N-terminal residues (amino acids at the amino terminus of the peptide chain) of light and heavy chains, and the remainder of the molecules are common (or constant) within a given isotype.

Some immunoglobulins, notably IgM (immune macroglobulin), exist as polymers of the basic four-chain unit. The IgM molecule normally consists of a cyclic pentamer of $L^2\mu^2$ subunits. This is the first immunoglobulin to appear in ontogeny, and the first to appear in phylogeny. Immunoglobulins of cyclostomes, elasmobranchs (sharks and rays), and many teleost fishes consist only of IgM polymers (5-mers in elasmobranchs, 4-mers in teleosts) and in some case IgM subunits (μ^2L^2). Immunoglobulins possessing heavy chains distinct from the μ-like heavy chains of those groups are present in some lungfish (Dipnoi) and in anuran amphibians (frogs and toads). Dipnoi have a low-molecular-weight non-IgM immunoglobulin (termed IgN) that possesses a heavy chain with a molecular weight of approximately 36,000. The μ chain has a weight of approximately 70,000, and the γ chain approximately 50,000. Anuran amphibians show a classical progression of antibody activity from high-molecular-weight molecules (IgM) to molecules of approximate weight 180,000, which is apparent with time following immunization. The low-molecular-weight immunoglobulin is not IgM, and it does not correspond directly to IgG or other of the non-IgM immunoglobulins of mammals. Amino acid sequence data are required to allow definitive conclusions regarding its direct homology to mammalian classes, but the data suggest that the heavy chain of the low-molecular-weight immunoglobulins of anuran amphibians represents a parallel duplication of a heavy-chain gene distinct from that which generated the gamma chain of mammals. Birds possess IgM and IgA immunoglobulins, but also possess a non-Ig immunoglobulin similar to that of amphibians as their major immunoglobulin class. This immunoglobulin has been termed IgY. IgG immunoglobulins containing gamma chains clearly homologous to those of humans and of true mammals are found only within the three subclasses of living mammals: eutherians, metatherians (marsupials), and monotremes (for example, the echidna).

Duplication of constant-region genes appears to be an essential feature of immunoglobulin evolution both on the large scale of emergence of vertebrate classes and on the smaller scale of speciation where distinct subclasses still arise from the major classes which might have emerged early in vertebrate evolution. The constant regions consist of domains or internal homology units of approximately 100 amino acids and contain an internal disulfide bond. These domains are specified by individual gene segments or exons occurring within the heavy-chain chromosome. For example, the μ chain consists of five domains, one variable and four constant, each of which is encoded by a separate exon. The emergence and evolutionary variation of the constant regions in immunoglobulin has resulted from duplication or deletion of gene segments corresponding to the domains of the actual proteins. Selective forces have acted upon the proteins produced because each domain carries out a particular effector function, such as fixation of complement or binding to cell surfaces. *See* EXON; IMMUNOGENETICS.

Variable-region diversity in lower vertebrates. Although all vertebrates can produce recognizable immunoglobulin, and sharks and higher vertebrates possess immunoglobulin clearly similar to that of mammals, the degree of diversity within the V-region compartment of lower vertebrates is probably less than that of mammals. Sharks, for example, can make specific antibodies to a wide variety of antigens, including small organic molecules (haptens), viruses, and foreign cells, but the affinity of these antibodies is low and does not increase with the duration of immunization. Although the size of the antibody can shift from IgM pentamer to IgM monomer with prolonged immunization, only the IgM class is represented and the affinity does not increase. By contrast, immunization of a common mammal, such as the rabbit, with a hapten usually generates first a primary response which is predominantly IgM with low affinity in the combining site. However, with prolonged immunization, the predominance of antibody activity is shifted to the IgG class and the affinity of the antibody for the hapten may increase by three or four orders of magnitude. Failure of affinity maturation is also a property of IgM antibodies of higher vertebrates and may reflect a lack of somatic mutation in the V_h genes that become associated with the μ constant regions. In addition, analyses, such as isoelectric focusing (which provides a means of counting the number of distinct antibodies on the basis of charge) and amino acid sequence analysis, indicate that in the shark and in the anuran amphibian *Xenopus*, the number of antibodies capable of responding to a hapten, such as dinitrophenol, is much less than that of mammals. The evidence suggests that the degree of heterogeneity of amphibian and fish antibodies is less than that of mammals. Although the basic event underlying immune recognition must be ancient and possibly preceded the origin of vertebrates, subsequent developments for increasing diversity, such as somatic mutation, apparently do not occur in lower vertebrates.

The primal event in the emergence of the classic immune system was the generation of V-region genes, probably by tandem duplication of a precursor and the joining of one of these genes to a constant-region gene in order to allow formation of an immunoglobulin heavy or light chain. It is possible that V-region genes and their products preceded constant regions which are necessary for effector functions because specific alloreactive lymphocytes, comparable to T cells, occur in species, such as tunicates, which lack the capacity to form circulating antibodies. Substantial evolutionary conservation has been found by comparing the sequences of heavy-chain variable-region genes of a reptilian species, the caiman, with those of humans and mouse. It is interesting to note that the caiman gene sequence clearly fits into the third heavy-chain variable-region subgroup of humans. This indicates that the generation of these subgroups preceded the events of speciation which led to the demarcation between reptiles and mammals.

Although the precise nature of the precursors of the specific elements of the immune system in evolution remains to be determined, the genetic and cellular events which lead to the capacity for specific immune recognition, diversification, and reactivity occurred early in vertebrate evolution. V- and C-region genes and a mechanism for joining them must have emerged prior to the emergence of true vertebrates, although somatic diversification mechanisms resulting in amplified repertoires and affinity maturation may not have arisen until relatively late in vertebrate evolution corresponding to the phylogenetic levels of birds and mammals. The specific immune elements were grafted onto a primitive and universal mechanism involving phagocytic cells and lectins. *See* IMMUNITY; IMMUNOLOGY. John J. Marchalonis

Bibliography. N. Cohen and M. M. Sigel (eds.), *The Reticuloendothelial System*, vol. 3: *Phylogeny and Ontogeny*, 1982; L. Du Pasquier, Antibody diversity in lower vertebrates—Why is it so restricted?, *Nature*, 296:311–313, 1982; M. Feldmann and N. A. Mitchison (eds.), *Immune Regulation*, 1985.

Immunology

The division of biological science concerned with the native or acquired response of complex living organisms to the intrusion of other organisms or foreign substances. The immune system allows the host organism to distinguish between self and nonself and to respond to a target (termed an antigen). An immunological response is specific and can be genetically determined or learned, that is, developed as a consequence of exposure. Examples of immune responses are neutralization of toxic protein molecules, killing of infecting microorganisms, rejection of foreign tissue grafts, and memory by the protective effects of vaccination. The last, discovered by E. Jenner toward the end of the eighteenth century, was the primary immunological finding.

However, it was not until the germ theory of infectious disease was established that the full implication of immunology was realized. First came the recognition that certain bacteria caused corresponding diseases; for example, *Salmonella typhi* caused typhoid fever and not whooping cough. Second came the recognition that it was a specific resistance to that bacterium or its toxins that prevented recurrence of the same disease. Third came the discovery that after recovery from an infectious disease, protective substances called antibodies could be found in the blood of animals and humans that would react with bacteria and their products in the living body and in the test tube. A host of serological reactions were devised to detect the presence of specific antibodies in serum. Antigens, such as bacteria and their products, triggered the production of antibodies and indeed all kinds of chemical and biological molecules.

Immune system. Lymphocytes are the principal cells of the immune system. There are two chief groups of lymphocytes, the thymic or T lymphocytes (T cells) and the B cells. T cells participate in most immunological responses. The thymus matures helper T cells that assist immune recognition and other (cytotoxic) T cells that are capable of suppressing immune reactions as well as of attacking other cells. T cells recognize antigen by its association with cell membrane–bound receptors. B lymphocytes, after stimulation with antigen molecules, and usually under the influence of factors released by helper T cells, proliferate and later secrete antibody molecules. B cells recognize antigen by means of cell membrane–bound antibodies. The bone marrow has been identified as the ultimate repository and source of a small number (perhaps a thousand) of stem cells that give rise to all the varieties of lymphocytes. *See* CELLULAR IMMUNOLOGY; IMMUNOGLOBULIN.

The mechanism enabling the recognition of self and producing the enormous variety of molecules by which nonself is recognized and countered has been found to reside in the genetic code. A basic code forms the foundation: one set of code corresponds to the immunoglobulins and is expressed in the B cells; another set corresponds to the membrane-fixed receptors on the T cells. *See* IMMUNOGENETICS.

After a foreign antigenic substance, such as a microorganism, protein, or carbohydrate, is recognized as being nonself, it is degraded by phagocytic cells. A portion of the antigen becomes associated in the cell with histocompatibility antigen molecules (antigen processing) and is routed to the cell surface, where it is presented for detection by lymphocytes. The specialized cells with this function are part of the lymphoid reticuloendothelial system and include dendritic cells and macrophages as well as lymphocytes. Lymphoid cells which recognize antigens may multiply in number as well as generate long-lived descendants that provide for memory. The latter allows an accelerated response in the event of subsequent exposure to the same antigen.

A variety of signaling molecules, such as the interleukins, assist the communication between cells. Effector cells which implement some immune reactions or participate in the removal of immunogenic materials include phagocytes, cytotoxic lymphocytes, mast cells, and natural killer cells. The effector mechanisms that kill and remove targeted agents also include a complex of blood proteins known collectively as complement, which has numerous effects of its own, as well as being able to interact with effector cells. *See* COMPLEMENT.

The action of these effector mechanisms, however, has come to be recognized as being not always protective or conferring immunity, but sometimes becoming grossly exaggerated or inappropriate, or capable of turning upon the host in a destructive fashion that causes disease. Exaggerated or inappropriate responses are typically associated with the production of certain antibodies to inhalants, contactants, foods and other normally innocuous products, as well as therapeutic materials or drugs. These responses are classified as allergies. Illnesses associated with a misguided response of the immune system that is directed against the self and results from a breakdown in the normal immunological tolerance of, or unresponsiveness to, self antigens are termed autoimmune, and include systemic lupus erythematosus and rheumatoid arthritis. The mechanisms responsible for these disorders are unknown but probably include the intervention of factors such as viruses that either modify or naturally resemble self molecules. Subsequently, the immune response, in seeking out what is foreign, proceeds to attack the self. Knowledge that would permit control of these aberrant responses would also illuminate the mechanisms involved in the rejection of transplants, thereby enhancing the exploitation of such procedures. *See* ALLERGY; AUTOIMMUNITY.

Related disciplines. Immunology is an eclectic science drawing on many other branches of knowledge. The comprehensive study of infection and resistance may be divided arbitrarily into data relating to the pathogen, the host, and their interactions during infection.

Important information on the host includes its species and the anatomical, physical, genetic, and nutritional factors that may influence infection and resistance. Host-parasite interactions include some classification of disease types, because one microbial species (such as the hemolytic streptococcus, *Streptococcus pyogenes*) may cause multiple diseases (septicemia, boils, scarlet fever, pneumonia, and rheumatic fever), with diverse immunological consequences. Many manifestations of pathology (for example, fever and inflammation) influence the reactivity of both host and parasite. Of prime concern to immunology are the cellular factors (lymphocyte accessory cells) and the humoral factors (antibodies and complement) that are significant aids in host resistance. *See* BIOASSAY; COMPLEMENT; FEVER; PATHOLOGY.

Immunology is also concerned with assaying the immune status of the host through a variety of serological procedures, and in devising methods of increasing host resistance through prophylactic vaccination. There has also been much important investigation of induced resistance and tolerance to transplants of skin and organs, including tumors. *See* ACQUIRED IMMUNOLOGICAL TOLERANCE; BLOOD GROUPS; IMMUNITY; IMMUNOASSAY; ISOANTIGEN; PHAGOCYTOSIS; PRECIPITIN; SEROLOGY; TRANSPLANTATION BIOLOGY; VACCINATION.

Alexander Baumgarten; Margaret J. Polley

Bibliography. D. J. Bibel, *Milestones in Immunology: A Historical Exploration*, 1988; F. H. Garrison, *An Introduction to the History of Medicine*, 4th ed., 1929; W. E. Paul, *Fundamental Immunology*, 4th ed., 1998; I. M. Roitt, *Essential Immunology*, 9th ed., 1997.

Immunonephelometry

An application of nephelometry to the quantification of antigen or antibody. Owing to the antigenic property of many macromolecules, immunonephelometry is capable of very broad application. In addition to its use for the quantification of soluble materials, it has been applied to the assay of microparticulate substances. The method permits the assay of antigen concentrations as low as 1 microgram per milliliter. *See* ANTIBODY; ANTIGEN; LIGHT-SCATTERING TECHNIQUES.

Method. The technique depends on the light-scattering properties of microparticles. An intense beam of collimated white light or a monochromatic beam (such as a laser) is passed through a solution. Scattering of the light is measured by a suitable detector (such as a photomultiplier tube) placed at an angle to the beam of light. As the number of particles which scatter light increases, so does the intensity of the detected light. A mixture of soluble antigen and antibody initially scatters no light. The initial antigen-antibody complexes are macromolecular in size and do not scatter light. However, such complexes have the ability to aggregate. The size of the aggregates increases with time until they do scatter light. This process takes from several minutes to hours before it attains maximum, depending on various properties of the reactants. If other aspects are kept constant, including, for example, the concentration of one reactant, then both the rate at which the scatter of light increases and its maximal value increase with the concentration of the other reactant. Either the increase in the rate of scatter or its maximal value can be calibrated for various concentrations of a reactant to yield a concentration-response calibration curve. Unknown concentrations of that reactant can then be quantified by measuring the response and comparing it with the calibration curve.

Although the maximal (equilibrium) values are maintained for a time (minutes to hours), they eventually decrease. This happens because, as the size of the aggregates continues to increase, their number and total reflective surface area decrease. This process eventually results in the formation of conventional antigen-antibody precipitates. Accordingly, rate immunonephelometry, which measures the rate of increase in light scatter, may be preferable to a technique which depends on the determination of "equilibrium" or maximal light scatter, because the former method is less sensitive to variation in the absolute duration of the reaction or precise geometry of the container and can generally be performed much more rapidly.

Reactants. The process of aggregation can be markedly accelerated by polyionic substances such as polyethylene glycol, which are therefore commonly added to the solution of reactants. It should be noted that the concentration of the reactants must be adjusted within certain limits, because antigen-antibody complexes are soluble in high relative concentrations of either reactant. The actual limits depend chiefly on the behavior of the antibody, so that some antisera are better than others for immunonephelometric use. The behavior of each antiserum must be individually assessed. The specificity of the immunonephelometric measurement of antigen depends entirely on the monospecificity of the antiserum, that is, its ability to react with only one antigen in a mixture of many.

Inhibition immunonephelometry. A variant of the immunonephelometric technique, inhibition immunonephelometry, has been used for the assay of substances of low molecular weight (haptens). The assay uses a constant amount of antibody and a predetermined quantity of the same reactant as that assayed, chemically coupled to a macromolecular carrier. The mixture yields a high nephelometric value of known magnitude. An introduction of the low molecular substance in the tested sample reduces the known magnitude of the nephelometric value in proportion to the concentration of the substance added in the tested sample. The extent of the reduction is used to determine the concentration of the assayed material in the sample by reference to a calibration curve. *See* IMMUNOASSAY. Alexander Baumgarten

Bibliography. C. L. Cambiaso et al., *J. Immunol. Meth.*, 5:293, 1974; F. W. Putnam (ed.), *The Plasma Proteins: Structure, Function and Genetic Control*, 5 vols., 2d ed., 1987; R. F. Ritchie, in F. W. Putnam (ed.), *The Plasma Proteins*, p. 375, 1975.

Immunopathology

The manifestations of conditions in which normally protective antibody or T cell immune mechanisms are overactive or expressed against self antigens, resulting in damage to cells or organs; also the study of these conditions. *See* ANTIGEN; ANTIBODY; IMMUNOLOGY.

The function of the immune system is to recognize molecular structures that are not normally part of the host's body, and to arrange for their inactivation and removal. Foreign structures may be pathogens; therefore this basic role of immunity is vital to survival. The innate immune system recognizes a few structures called pathogen-associated molecular patterns (PAMP), while the adaptive immune system can respond to billions of molecular configurations. When immunity is activated, the events are intense: cells divide rapidly and release powerful proinflammatory chemokines and cytokines. Such a system needs two types of control: it must avoid responding to normal, "self" structures, and it must regulate the intensity of its reaction so that tissues and organs are not seriously damaged as "innocent bystanders." When the controls work well, the functioning of the immune system is almost unapparent, although it is essential for survival in a world full of pathogens. When the controls fail, we observe immunopathology. *See* IMMUNITY; PATHOGEN; PATHOLOGY.

At least 200 conditions have been described in humans and animals in which the immune system is involved in tissue damage. This array would be bewildering had not P. G. H. Gell and R. R. A. Coombs in 1963 shown that almost all examples of immunopathology (excluding immunodeficiency and tumors of the immune system) could be assigned to one of four categories, based on the underlying immunological mechanism. Although modifications and additions to these categories have been suggested over the years, they remain the fundamental way in which clinicians look at immunopathology. *See* CELLULAR IMMUNOLOGY; CLINICAL IMMUNOLOGY.

Type I immunopathology. Type I immunopathology is also called allergy or immediate hypersensitivity. The mechanism involves excessive production of the immunoglobulin E (IgE) class of antibody, usually in response to harmless environmental antigens. IgE seems to play its major normal role in resistance to infestations by parasites such as intestinal worms and protozoans such as *Giardia*, *Cryptosporidium*, *Trypanosoma*, and *Schistosoma*. In type I immunopathology the antigen may be a harmless substance entering through the mucous membranes, such as a food or pollen. Why some people react this way to these allergens is unknown, though there is a strong genetic predisposition to it. Allergic people make strong IgE responses, and the antibody adheres to IgE receptors on the surface of mast cells (connective-tissue cells with numerous granules rich in histamine and heparin) in the tissues, especially those of the skin, eyes, nose and throat, and gastrointestinal tract. When further allergen enters the body, the allergen binds to the specific IgE that was generated against it, and crosslinks it on the mast cell surface. This constitutes a signal to the mast cell to rapidly release granules full of preformed histamine, and to begin synthesis of other inflammatory mediators, including prostaglandins and leukotrienes (thus antagonists of these classes of mediators are mainstays of allergy treatment). Mast cells also release substances that are chemotactic for (attract) blood eosinophils (granular leukocytes, or white blood cells). In response to protozoans, for example, these white blood cells are essential, as they contain powerful agents that can destroy them, something the more common neutrophils cannot achieve. In allergy, the eosinophil infiltration may help diagnosis, but if anything it worsens the inflammatory condition. *See* ALLERGY; HYPERSENSITIVITY; IMMUNOGLOBULIN.

Allergic syndromes are categorized on the basis of the organ most affected. Histamine in general causes contraction of smooth muscle in the respiratory and gastrointestinal tracts, and relaxation of smooth muscle in blood vessels, which leads to vessel leakiness. In the skin, release of histamine produces the intensely itchy hive, or wheal-and-flare reaction. In the intestine histamine speeds up peristalsis, which can result in diarrhea and vomiting. Constriction of smooth muscle in the respiratory tract leads to sinus blockage, wheezing, or asthma, as well as the runny eyes and nose that are called (incorrectly, since hay is rarely an allergen) hay fever. If there is sudden systemic histamine release, as might follow ingestion of certain foods or an insect sting, there can be a fall in blood pressure so pronounced as to result in anaphylactic shock and death. Agents that constrict blood vessels and dilate airways, such as epinephrine, can be life-saving. *See* ASTHMA; HISTAMINE.

Allergies are detectable by intradermal skin tests with the suspect agent; a wheal-and-flare that develops within 15 min (hence, it is also called immediate hypersensitivity) helps in making the diagnosis. Lab tests for elevated levels of specific IgE antibody to many known allergens, called RASTs (radioallergosorbent tests), are available. Total blood IgE may be elevated as well. *See* SKIN TEST.

Type II immunopathology. In type II immunopathology, autoimmune antibodies are made that cause damage not to pathogens but to normal body cells. T and B lymphocytes are exposed during their maturation to a wide variety of self molecules. Any lymphocyte whose receptors happen to bind self with high affinity are deleted from the repertoire, since they would be potentially autoimmune. This phenomenon is known as negative selection, and results in us being "tolerant" to our own tissues. In rare individuals there are vital genetic defects in the process and severe autoimmunity results. In most of the more common autoimmune diseases such as rheumatoid arthritis, systemic lupus erythematosus (SLE), and Graves' disease, multiple genes and environmental influences have been implicated. *See* AUTOIMMUNITY.

A case in which the cause is known is rheumatic heart disease. When certain individuals develop an infection with *Streptococcus*, the antibodies made against the bacteria cross-react with a structure expressed on the lining of the heart, especially around the valves. The antibodies bind to the heart, complement is activated, neutrophils are attracted, and damage to the tissues results. All of this represents a normal immune mechanism misdirected against a body structure. In developed countries, children are treated aggressively with antibiotics for streptococcal infections, leading to a steep decline in rheumatic heart disease, but it is still common in the developing world. *See* HEART DISORDERS; RHEUMATIC FEVER.

The expression of autoimmunity may be focal or general. In autoimmune hemolytic anemia the red blood cells throughout the body are the target, and in myasthenia gravis it is the acetylcholine receptor, involved in neural stimulation of muscles. Type 1 diabetics make antibodies that damage the insulin-producing beta cells in the pancreas. An unusual mechanism is seen in Graves' disease, an autoimmune hyperthyroid condition. Here the antibodies are directed against the thyroid-stimulating hormone receptor, but instead of causing cell destruction they act as powerful stimulators at the receptor, driving the thyroid to overproduce its hormones. *See* ANEMIA; DIABETES; MYASTHENIA GRAVIS; THYROID GLAND DISORDERS.

Treatment involves immunosuppressive drugs and powerful anti-inflammatory agents. Side effects are thus common, including increased susceptibility to infection. Sophisticated new drugs and biological response modifiers are coming onto the market with promising results.

Type III immunopathology. Type III immunopathology is also called immune complex disease, and again involves a normal mechanism that becomes dysregulated. Because antibodies have two combining sites for antigen, and antigens may have many antigenic determinants, antibodies can form immune complexes, three-dimensional lattices of many antigens and antibodies strung together. Large complexes are readily cleared from the blood by the phagocytic cells. Smaller ones may get trapped in the basement membranes, a sort of molecular mat that underlies the cells lining small blood vessels. The trapped complexes activate complement and an inflammatory response ensues. This is most common where there is a net outflow of fluid from blood to tissues: in the joints, pleura, choroid plexus of the brain, skin, and most importantly the kidneys, in which glomerulonephritis (inflammation of the kidney) can develop. *See* ANTIGEN-ANTIBODY REACTION; IMMUNE COMPLEX DISEASE.

The antigens involved may be external to the body or they may be self. The condition was first described in 1905 in children who, in the preantibiotic era, had received large doses of serum from diphtheria-immunized horses. Although the diphtheria was controlled, in about a 7 to 10 days the patients would develop rash, joint pain, fever, and signs of glomerulonephritis. They had made antibodies against the foreign horse serum proteins, and these were appearing in the blood at a time when there were still considerable amounts of the horse proteins circulating. Immune complexes formed and deposited in the joints, skin, and kidneys. This condition is still seen in people who receive foreign sera, such as for the treatment of snake bite, and is called serum sickness.

If a person's immune system mistakenly forms antibody against a protein normally found in the circulation, chronic type III immunopathology may result. Thus people with systemic lupus erythematosus make antibodies to DNA, of which there is always a small amount in the blood from the normal turnover of cells. In rheumatoid arthritis, antibody is actually made against antibody.

Type III immunopathology is treated by decreasing the concentration of antigen if possible, and by the same anti-inflammatory and suppressive regimens used for type II.

Type IV immunopathology. This is the only one of the four types that involves damage to body tissues by T cells. It can be the result of an excessive response or true autoimmunity. One of the classic type IV conditions is due to poison ivy. When a person is first exposed to the resin of the plant, one of its constituents, urushiol, penetrates intact skin and associates with antigen-presenting cells which carry it to the local lymph nodes. Th1 (T helper 1) cells respond, proliferate, and become active. By the time (a week or two) these T cells are circulating around the body, the resin is usually gone from the skin, and nothing is observed, so the victim is unaware that he or she has become "sensitized." However, a subsequent exposure means the T cells are already there and they do what they are designed to: call up a powerful, macrophage-type, inflammatory response. The skin is damaged, although it is an innocent bystander. This condition is called delayed hypersensitivity or contact hypersensitivity, and can be seen with many antigens, including nickel, cosmetics and dyes, and latex. *See* IMMUNOLOGICAL ONTOGENY; POISON IVY.

Other T cell–mediated conditions involve true autoimmunity. For example, in multiple sclerosis, clones of T cells attack antigens derived from the myelin sheaths of central nervous system neurons. What causes this—possibly cross reactions between myelin and an unknown environmental agent—remains unknown.

Many of the autoimmune diseases involve more than one immunopathologic mechanism at the same time: people with rheumatoid arthritis have both antibodies against antibody and T cells against joint collagen, for example, and both antibodies and T cells are implicated in type 1 diabetes. Because the immune system must be able to respond to a near-infinity of antigens, it sometimes becomes confused; and because it is so complex, there is still much research and clinical testing to be done. J. John Cohen

Bibliography. H. Chapel et al. (eds.), *Essentials of Clinical Immunology*, 5th ed., Blackwell, London, 2006; C. Janeway, Jr., et al. (eds.), *Immunobiology*, 6th ed., Garland Science, New York, 2005; D. Male et al. (eds.), *Immunology*, 7th ed., Mosby Elsevier, Philadelphia, 2006.

Immunosuppression

The natural or induced active suppression of the immune response, as contrasted with deficiency or absence of components of the immune system. Like many other complex biological processes, the immune response is controlled by a series of regulatory factors. A variety of suppressor cells play a role in essentially all of the known immunoregulatory mechanisms, such as maintenance of immunological tolerance; limitation of antibody response to antigens of both thymic-dependent and thymic-independent types, as well as to antigens that stimulate reaginic antibody (antibodies involved in allergic reactions); genetic control of the immune response; idiotype suppression; control of contact and delayed hypersensitivity; and antigenic competition. All of the major cell types [macrophages (monocytes), bone marrow–derived lymphocytes that normally mature into antibody-producing cells (B cells), and especially thymus-derived lymphocytes (T cells)] involved in the positive side of cellular interactions required for an immune response have also been found capable of functioning as suppressors in different regulatory systems. *See* IMMUNOLOGICAL DEFICIENCY.

Suppressor cells. Some suppressor functions are antigen- or carrier-specific. (A carrier is a molecule that can be chemically bound to another small molecule, called a hapten, in such a way that the combination induces an immune response that the hapten alone would not induce.) Others may not be carrier-specific, but may be specific for the type of response, such as immunoglobulin production but not delayed hypersensitivity. In the case of immunoglobulin production, the suppressor T cell may regulate the production of all immunoglobulin classes (IgG, IgM, IgA, and so on), a single class of immunoglobulins, or molecules that bind only a given antigen. Other suppressors may affect only cellular immunity and not humoral immunity. *See* CELLULAR IMMUNOLOGY; IMMUNOGLOBULIN.

Regulation of normal immune response. Suppressor cells are critical in the regulation of the normal immune response. Immunological tolerance refers to the ability of an individual's immune system to distinguish between its own and foreign antigens and to mount a response only to foreign antigens. A major role has been established for suppressor T lymphocytes in this phenomenon, although a complex array of other mechanisms that are not mutually exclusive are also implicated.

Suppressor cells play a role in regulating the magnitude and duration of the specific antibody response to an antigenic challenge. This regulatory role was first demonstrated with antigens that are thymic-dependent, requiring helper T cells as well as B cells for an optimal response. Subsequently it was shown that antigens that are relatively thymic-independent are also regulated in both a positive and negative manner by different sets of T cells.

Immune-response genes determine the capacity to mount an immune response to certain well-defined antigens in a number of species. There is evidence that in some cases suppressor cells play a major part in inducing the nonresponder status, but that they cannot account for it in most situations. Other mechanisms must be active as well.

A mitogen is a substance, usually plant-derived, that induces cells to divide. The induced mitosis is accompanied by changes identifiable under the microscope from a mature-appearing cell to a "blast" or immature cell. Specific mitogens have been found to activate T cells in antigen-nonspecific ways. Concanavalin-A is especially effective in stimulating the suppressor T-cell system that can suppress the generation of lymphocytes which are cytotoxic to allogeneic cells (cells from another animal of the same species), the mixed lymphocyte reaction, mitogen-induced blastic transformation, and antibody and immunoglobulin production.

Reagin, or IgE, is the class of immunoglobulin that mediates allergic reactions such as asthma and urticaria. The reaginic antibody response depends heavily on nonspecific cooperator T cells and specific helper T cells as well as the B cells that produce the antibody. In a negative direction, IgG-blocking antibodies regulate the response, but antigen-specific and antigen-nonspecific suppressor T cells also play a critical role in regulating this response. *See* ALLERGY.

Allotype refers to inherited variations in the genes coding for certain amino acid sequences in the constant region of the immunoglobulin molecules. An animal can be treated in a manner that prevents production of an inherited allotype. In some circumstances, suppressor cells maintain this suppression. Cells capable of producing the allotype can be demonstrated if the suppressors are removed. Idiotype refers to the specific structure of the hypervariable regions that bind antigen, and in a similar manner it has been shown that suppressor cells maintain suppression of idiotype production. *See* IMMUNOGENETICS.

Antigenic competition refers to the observation that immunization with one antigen interferes in a nonspecific manner with the immune response to a subsequently administered second antigen. This phenomenon is short-lived and is mediated by macrophages rather than by T lymphocytes.

Mechanisms of suppressor cell action. T cells are the major cells involved in immunosuppression, although activated phagocytic mononuclear cells are also significant as nonspecific suppressors in many systems. Helper T cells and suppressor T cells are different cell populations that are distinguished to a considerable extent by surface antigens that react with monoclonal antibodies (such as T_4 and T_8) or receptors for specific substances such as histamine. *See* MONOCLONAL ANTIBODIES.

The suppressor T cell emerges from the thymus as an inactive precursor and matures into an effector of suppression after antigen stimulation. The developmental process does not involve antigen stimulation alone, but also requires the interaction of interleukin I and interleukin II with their receptors on T cells. Interleukins I and II are antigen-nonspecific soluble factors produced by monocytes and T cells, respectively, that are involved in the expansion of the numbers of T lymphocytes including suppressor cells as well as helper and cytotoxic T-cell populations.

No single model explains the entire array of cellular suppressor phenomena. In different systems, other T cells, macrophages, or even B cells may be the immediate targets of the suppressor cells and their secretions. Some suppression requires direct cell–cell interaction, whereas other suppression may be mediated by suppressor lymphokines (soluble factors produced by lymphocytes). Both antigen-specific and antigen-nonspecific factors are known, and they may be secreted to act upon other cells, or especially in the case of antigen-specific factors, they may be integral parts of the cell membrane. The soluble immune-response suppressor factor, produced by activated T cells and then activated by monocytes, inhibits B-cell proliferation and immunoglobulin production in response to antigens. Macrophages also secrete suppressor factors, including prostaglandins that act on T cells and other soluble factors that are B-cell–specific.

Disorders of suppressor cell regulation. There is a variety of disorders of immunoglobulin production in humans. In many cases these involve intrinsic defects in the bone marrow stem cells that normally mature into immunoglobulin-producing plasma cells. However, some individuals have been found to have B lymphocytes capable of maturing into cells that secrete immunoglobulins when separated from their own suppressor T cells. The T cells from such individuals have been demonstrated to suppress the production of immunoglobulins by B cells of normal humans. In animal models of immunoglobulin deficiency, the development of suppressor T cells may occur as a secondary event that maintains suppression of immunoglobulin production even when the primary event is a defect in the B-cell line. A subgroup of individuals with deficiency of only the IgA class of immunoglobulins has suppressor T cells that prevent B cells of normal humans from producing IgA.

Defects in cell-mediated immunity occur in individuals who are infected with various fungal organisms. Suppressor T cells have been implicated, although it is not clear whether the appearance of suppressor cells is the initial event allowing development of the fungal infection or whether they develop secondarily after infection. Those individuals found to have suppressor T cells are at high risk for dissemination of the fungal infection and relapse following therapy.

Although probably only one of many mechanisms, suppressor cells interfere with the host tumor-growth–inhibiting immune response to the foreign tumor-specific transplantation antigens that occur on malignant cells, thus allowing the tumors to progress. In cases of multiple myeloma, there is a significant reduction in polyclonal immunoglobulin synthesis, and serum levels of polyclonal immunoglobulins are low, though there is a high production of the monoclonal immunoglobulin by the malignant plasma cells. Affected individuals are particularly susceptible to infection by pathogenic encapsulated bacteria. One mechanism has been shown to be a blockage of polyclonal lymphocyte maturation by host regulatory macrophages. The acquired immunodeficiency principally affecting cell-mediated immunity in Hodgkin's disease is also caused by macrophage suppressors. Hypogammaglobulinemia associated with benign thymoma has been demonstrated to be associated with circulating suppressor T cells.

Both animal and human studies indicate a major role for both an activation of immunoglobulin-producing B cells as well as the absence or reduced numbers or function of suppressor T cells in autoimmune disorders such as Coombs-positive hemolytic anemia, systemic lupus erythematosus, rheumatoid disorders, and thyroid disorders in which antithyroid antibodies appear in the serum. *See* AUTOIMMUNITY.

Katharine S. Waldmann; Thomas A. Waldmann

Immunosuppressors. Suppression of the immune response may be specific to a particular antigen or may be a response to a wide range of antigens encountered. The whole immune response may be depressed, or a particular population of immunologically active lymphocytes may be selectively affected. In some cases, the effect may be preferentially on T cells rather than B cells. If B cells are affected, it may be on a specific subclass of antibody-producing cells. Antigen-specific immunosuppression may be the result of deletion or suppression of a particular clone of antigen-specific cells, or the result of enhanced regulation of the immune response by antigen-specific suppressor cells. It can also be the result of increased production of anti-idiotypic antibody.

Nonspecific suppression of the immune response occurs in a number of rare immunological deficiency diseases of childhood. These include severe combined immune deficiency syndrome (SCID), in which both T lymphocytes and B lymphocytes are affected; deficiencies of development of the third and fourth branchial arches, affecting the thymus and thus T cells only; or deficiencies of B-cell development, resulting in defects in antibody synthesis. Other conditions are the IgA deficiency associated with ataxia telangiectasia and the variable immunodeficiency states. Acquired deficiency states affecting mainly T-cell function occur in states of malnutrition and in the presence of tumors, particularly those of the lymphoreticular system both in Hodgkin's disease and non–Hodgkin's type lymphoma. Acquired deficiences may also occur secondary to a number of infectious diseases, particularly those of viral origin, but including malaria and typhoid fever. The acquired immune deficiency syndrome (AIDS) is probably of similar origin; its manifestations are similar although more severe and more deramatic. *See* ACQUIRED IMMUNE DEFICIENCY SYNDROME (AIDS); IMMUNE COMPLEX DISEASE.

The effect of total body irradiation on the lymphoreticular system also produces an acquired immunodeficiency. Irradiation with x-rays or gamma rays has a direct effect on all rapidly dividing cell populations, though B lymphocytes and T lymphocytes differ in their susceptibility to irradiation. Circulating antibody production may be eliminated by as low as 200–300 roentgens of total body irradiation, but it takes 400–800 R to affect cell-mediated immunity. Depression of T-cell–mediated immunity results in increased susceptibility to opportunistic infections.

Drug-induced immunosuppression. There are a number of compounds capable of suppressing the immune response. The main stimulus for studies designed to identify these substances has been to devise means for controlling organ graft rejection. However, there has also been considerable activity in looking for compounds that will suppress the immune response and reduce the inflammatory process in experimental models of rheumatoid arthritis.

For a number of years it has been known that irradiation by x-rays or gamma rays suppressed the immune response, so it was logical to look first at compounds with a radiomimetic profile. These were mainly alkylating agents and derivatives of nitrogen mustards. Further research uncovered purine and pyrimidine analogs and antimetabolites such as the folic acid antagonists. Another group of agents with immunosuppressive properties is the fungus-derived antibiotics that suppress nucleic acid and protein synthesis in mammalian cells. Corticosteroids, antilymphocyte globulin (ALG), and cyclosporin-A (cyclosporine) complete the list.

In the treatment of recipients of organ graft transplants (kidney, liver, heart), there is an extremely fine balance between the dose which is necessary to inhibit graft rejection alone and that which will suppress the body's resistance to infection particularly with opportunistic organisms such as *Candida albicans, Staphylococcus albus*, *Pneumocystis carinii*, and *Cytomegalovirus*. A reduction in dose will allow development of a rejection crisis, whereas an increase may precipitate an overwhelming and uncontrollable infection with these and similar organisms. *See* TRANSPLANTATION BIOLOGY.

The ideal immunosuppressive drug should fulfill five main requirements: (1) There should be a wide margin of safety between a toxic and a therapeutic dose. (2) The drug should have a selective effect on lymphoid cells and not cause damage to the rest of the body. (3) If possible, this effect should be only on those cells which are involved in the specific immune process to be suppressed. (4) The drug should need to be administered for only a limited period until the immunological processes become familiar with the foreign antigen and begin to recognize it as part of "self." (5) The drug should be effective against immune processes once they have developed.

Alkylating agents. Not all alkylating agents employed in antitumor chemotherapy have a satisfactory therapeutic index when used as immunosuppressive agents, though they may produce immunosuppression as a side reaction in the treatment of cancer. The most powerful immunosuppressant in this group is cyclophosphamide, which is a phosphoric ester diamide. Cyclophosphamide is not active in the test tube and is broken down in the body in liver microsomes to its active components. These metabolites remain active in the body for no longer than 4 h following injection. As do other alkylating agents, cyclophosphamide produces single-strand breaks in deoxyribonucleic acid (DNA) following alkylation and also forms cross-links binding two strands of DNA. It acts therefore on cells during mitosis, and on more rapidly dividing than slowly dividing cells. Cyclophosphamide has been shown to be most active against the precursors of B cells and T suppressor cells, and thus if given before immunization can sometimes enhance the immune response. If given during the development of the immune response, it will block both antibody production and cell-mediated immunity, as it hits at T cells particularly during the proliferative phase of the immune response. Clinically, cyclophosphamide has been used mainly in the treatment of systemic lupus erythematosus, severe rheumatoid arthritis, and nephritis. It may also have a powerful nonspecific anti-inflammatory effect. The clinical use of this agent is limited by its extremely toxic side effects, which include hair loss and exfoliation of the bladder epithelium.

Thiopurines. The main thiopurines used as immunosuppressive agents are the guanine analogs 6-mercaptopurine and azathioprine. Azathioprine has the advantage of solubility over 6-mercaptopurine and is also considered to be therapeutically more active. Purine analogs compete with free purines for incorporation into nucleic acid and thus inhibit nucleic acid biosynthesis and protein synthesis. Azathioprine has been the standard immunosuppressant used in transplantation surgery since the early 1960s, and is usually supplemented by corticosteroid therapy. It may also be supplemented successfully by antilymphocytic globulin, which has proved particularly valuable in cardiac transplantation. Azathioprine has also been used in the treatment of severe rheumatoid arthritis and systemic lupus erythematosus. *See* NUCLEIC ACID.

Antimetabolites. The most commonly encountered antimetabolite with an immunosuppressive effect is methotrexate, a folic acid inhibitor. This blocks the reduction of dihydrofolinic acid to tetrahydrofolinic acid, which is a coenzyme used in the formation of thymidine from uracil deoxyriboside. Methotrexate is a potent immunosuppressive agent in a number of animal models.

Fungal metabolites. The actinomycins were used temporarily to supplement azathioprine in transplantation surgery. However, the most promising fungal metabolite is cyclosporin-A. Cyclosporin-A has a preferential effect on T lymphocytes and thus on cell-mediated immune processes. It is highly effective in transplantation surgery and can be used alone without azathioprine. It has a much wider margin of safety between the therapeutic and toxic levels and does not have a strong depressive effect on host resistance to infection.

Corticosteroids. There are considerable differences in the susceptibility of lymphoid tissue of various species to the action of corticosteroids. Whereas the hamster, mouse, rat, and rabbit are strongly steroid-sensitive and the action of these compounds is lympholytic, the ferret, monkey, guinea pig, and humans are relatively resistant. Moreover, whereas the action of corticosteroids in the mouse is primarily on lymphocytes, in the guinea pig a major effect is on cells of the mononuclear phagocyte series. The major use of corticosteroids in medicine is as nonspecific and anti-inflammatory agents, and it is likely that they act in this way when used to supplement azathioprine in transplantation surgery. *See* STEROID.

Antilymphocytic globulin. Antilymphocyte serum was first prepared by E. Metchnikoff in 1899, but it was not until the advent of transplantation surgery that it was shown to be such a potent immunosuppressive agent with a preferential action on cell-mediated immunity. Its action is directed mainly against T lymphocytes. The commercial preparation of a purified reagent for use in transplant surgery in humans has presented a number of problems, the most important of which is the difficulty in standardization. Different batches produced over the years have varied considerably in immunosuppressive activity and toxicity. The source of the antigen is critical. Thus thymus and thoracic duct lymphocytes seem to provide the best source. Although horses have been used in a number of commercial preparations, rabbits appear to provide a better serum. A further problem is that there is little correlation between tests of activity outside the body and efficacy in the body. Toxicity of preparations may be due to sensitization to heterologous serum proteins, anti–red cell antibodies, antiplatelet antibodies, and nephrotoxic antibodies (particularly antiglomerular basement membrane), all of which must be removed.

In addition to all these problems in preparation, purification of the antibody is necessary either by globulin fractionation or by absorbtion to lymphocytes followed by elution at low pH. However, there are a number of preparations available that have proved extremely valuable, especially in cardiac transplants in humans.

Drug-induced immunological tolerance. Specific immunological unresponsiveness can be enhanced by drug therapy, including cyclophosphamide, 6-mercaptopurine, and antilymphocytic globulin. The ease by which unresponsiveness can be induced in adult animals in which the numbers of both T and B cells are reduced confirms the impression from earlier work that this form of tolerance is produced by the action of antigen on immature cells. One might suppose that by reducing the number of available immature cells and most of the immmunologically mature cells, there develops a higher ratio of antigen to immature cells present. In this way the animal is pushed from the zone of immunization to that of high-zone tolerance, not by increasing the level of antigen, but by reducing the number of potentially active cells.

Drug-induced potentiation of immune response. The result of any immune response is a balance between the action of effector cells mediating the phenomenon and suppressor cells regulating the response. Anything that reduces the regulatory function of suppressor cells will functionally increase the immune response. As suppressor cells are derived from rapidly turning-over precursor cells, and effector cells of T-cell–mediated immunity are derived from slowly dividing precursors, it is possible preferentially to depress the action of suppressor cells without affecting effector cells. This may be done by the use of alkylating agents such as cyclophosphamide given before immunization. Cyclophosphamide used in this way can increase a normal cell-mediated immune response, reverse immunological tolerance caused by increased regulatory activity of suppressor cells, and even reverse antigenic competition. It is likely that the chemotherapeutic effect of alkylating agents which are used extensively in the treatment of cancer in humans is partially due to these agents modifying the biological response to the tumor, producing an immunopotentiating action. *See* CHEMOTHERAPY AND OTHER ANTINEOPLASTIC DRUGS; IMMUNITY; IMMUNOLOGY.

J. L. Turk

Bibliography. A. Allison, K. Lafferty and H. Fliri (eds.), *Immunosuppressive and Antiinflamatory Drugs*, 1993; J. F. Bach, *The Mode of Action of Immunosuppressive Agents*, 1986; G. Möller (ed.), Immunosuppressive agents, *Immunol. Rev.*, vol. 65, 1982; J. L. and D. Parker (eds.), *Drugs and Immune Responsiveness*, 1979; E. I. Reinherz and S. F. Schlossman, The characterization and function of human immunoregulatory T lymphocyte subsets, *Pharmacol. Rev.*, 34:17–22, 1982; R. S. Schwartz, *Progress in Clinical Immunology*, vol. 3, 1977; A. W. Thomson (ed.), *Cyclosporin: Mode of Action and Clinical Application*, 1989; J. L. Turk, *Delayed Hypersensitivity*, 3d ed., 1980.

Immunotherapy

The treatment of cancer by improving the ability of a tumor-bearing individual (the host) to reject the tumor immunologically. There are molecules on the surface of tumor cells, and perhaps in their interior, that are recognized as different from normal structures by the immune system and thus generate an immune response. These molecules, called tumor-associated antigens, have been clearly demonstrated with animal tumors and are strongly suggested with human tumors.

The two components of the immune response are cell-mediated and antibody-mediated immunity, which must work in concert to overcome tumor cells. One type of thymus-derived lymphocyte (also called a cytotoxic T cell) can destroy tumor cells directly, while another recruits other white blood cells, the macrophages, that do the killing. Natural killer cells and perhaps other white blood cells may also participate. Antibodies produced in reaction to a tumor probably work mostly by binding to white blood cells at one end and to tumor cells at the other, bringing the two cells into close proximity and facilitating destruction of the tumor cells. However, elements that normally regulate immunity, such as suppressor T cells, are stimulated excessively by the tumor, which leads to an immune response that is deficient and unable to reject the growing tumor. Thus the strategy of immunotherapy is to stimulate within or transfer to the tumor-bearing individual the appropriate antitumor elements while avoiding further stimulation of suppressor elements. *See* CELLULAR IMMUNOLOGY; IMMUNOLOGIC CYTOTOXICITY; IMMUNOSUPPRESSION.

There are four broad categories of immunotherapy: active, adoptive, passive, and restorative.

Active immunotherapy. This form of treatment attempts to stimulate the host's intrinsic immune response to the tumor, either nonspecifically or specifically.

Nonspecific. Nonspecific active immunotherapy utilizes materials that have no apparent antigenic relationship to the tumor, but have modulatory effects on the immune system, stimulating macrophages, lymphocytes, and natural killer cells. A large and rather heterogeneous group of agents have been studied, the best-known of which are microbial agents, such as bacillus Calmette-Guérin (BCG), and also pyran, interferon, and the vitamin A derivatives called retinoids.

Clinical trials with nonspecific active immunotherapy alone have largely failed to improve the survival of cancer patients, with only a few exceptions, such as in nodular lymphoma and in ovarian cancer. Interferon-alpha is a low-molecular-weight protein produced by leukocytes in response to viral infection. Initial studies with partially purified, extracted interferon-alpha reported several remissions in patients with multiple myeloma and nodular, poorly differentiated lymphocytic lymphoma, and a prolonged survival of patients with osteosarcoma which was treated prophylactically after amputation. Partial responses were also observed in late-stage breast cancer patients. Trials of recombinant interferon-alpha have noted significant responses of patients with renal cell (kidney) carcinoma, multiple myeloma (a bone marrow tumor), and Kaposi's sarcoma, a tumor often found in patients with acquired immune deficiency syndrome (AIDS). However, it is uncertain whether the nonspecific immunomodulatory effects of interferon were responsible, since interferon also has a direct antitumor effect. Studies on retinoids were prompted by epidemiological data showing a decreased incidence of lung and bladder cancer by the high intake of vitamin A. While retinoids do augment the immune response, like interferon they have direct effects on neoplastic cells. Retinoids also promote the maturation of preneoplastic cells toward normal development. *See* ACQUIRED IMMUNE DEFICIENCY SYNDROME (AIDS); VITAMIN A.

Specific. Specific active immunotherapy attempts to stimulate specific antitumor responses with tumor-associated antigens as the immunizing materials. Tumor cells from the same patient or from another with the same type of tumor, irradiated or altered in tissue cultures to make them more antigenic, have been used to try to augment specific immunity. Also, the potential usefulness of antigenic extracts, both soluble and particulate, has been investigated in such dissimilar diseases as malignant melanoma (malignant skin mole) and lung cancer. Rapid progress in identifying important tumor-associated antigens has been made with these and several other tumors, especially through the use of monoclonal antitumor antibodies. The use of extracts of tumor antigens for immunotherapy must be approached with great caution, since soluble tumor antigens can generate suppressor T cells when combined with a small amount of host antibodies. This could lead to enhanced tumor growth. The content of antigens in specific tumor vaccines, the form in which antigens should be presented to the patient (for example, particulate, together with adjuvant materials, and so on), and the schedule and route of administration are all critical questions. *See* MONOCLONAL ANTIBODIES.

Adoptive immunotherapy. This denotes the transfer of immunologically competent white blood cells or their precursors into the host. Bone marrow transplantation, while performed principally for the replacement of hematopoietic stem cells, can also be viewed as adoptive immunotherapy. Some success in the treatment of acute leukemia has been noted by allogeneic (second-party) bone marrow transplantation after sublethal irradiation of the patient. The results from syngeneic (twin) bone marrow transplants have been less encouraging, perhaps because the identical twin's lymphocytes do not recognize the leukemic cells as sufficiently different from self to reject them. It is, however, possible to grow T cells in tissue culture with the aid of T-cell growth factor (interleukin-2), stimulate (or restimulate) them with tumor antigens, and transfuse the cells into the patient. This has been effective therapy in mice. Transfusion of leukocytes armed with antitumor antibodies is another approach that appears feasible for the treatment of human cancers.

Restorative immunotherapy. This approach comprises the direct and indirect restoration of deficient immunological function through any means other than the direct transfer of cells. Repletion of competent effector-cell populations is one form of restorative immunotherapy. Thymic hormones, such as thymopoietin and thymosin, can convert T-cell precursors into helper T cells, a property shared by the antihelminthic agent Levamisole and several of its congeners. In a study of small-cell carcinoma of the lung, thymosin increased the median survival of patients who had sustained a complete response from chemotherapy. The subgroup that benefited most was the one with a deficient complement of T cells. Deficient functioning of T cells may be a problem in most patients with a tumor. Deficient percentages of T cells relative to other leukocytes have been noted in several types of tumor, the most noteworthy of which is Kaposi's sarcoma. Repletion of number or function is an important strategy which might by itself be curative in certain tumors. Antagonism of suppressor influences is another form of restorative therapy. Low doses of the antitumor drug cyclophosphamide selectively inhibit suppressor T cells. Inhibitors of prostaglandin synthesis, such as indomethacin, may be useful in overcoming other suppressor cells of the macrophage series, which act principally through prostaglandins. *See* EICOSANOIDS.

Passive immunotherapy. This means the transfer of antibodies to tumor-bearing recipients. This approach has been made feasible by the development of hybridoma technology, which now permits the production of large quantities of monoclonal antibodies specific for an antigenic determinant on tumor cells. Monoclonal antibodies made against tumor-specific antigens, if they in fact exist, or against tumor-associated antigens, which may be shared by certain normal cells, are of potential importance therapeutically. These antibodies probably work through the arming of white blood cells, especially macrophages, of the host. They may be most effective in treating dispersed tumors such as leukemia, and ovarian cancer in ascites (abnormal accumulation of serous fluid in the abdominal cavity). The potential for enhancement of tumor growth, as a result of the formation of antigen-antibody complexes and the excessive stimulation of suppressor cells, exists and must be carefully monitored. The development of human monoclonals should obviate some of the allergic reactions that have occurred with mouse antibodies, and promises still greater specificity of attack on tumor cells. *See* CANCER (MEDICINE); GENETIC ENGINEERING; IMMUNOLOGY.

June Kan-Mitchell; Malcolm S. Mitchell

Bibliography. P. Calabresi and P. S. Schein (eds.), *Medical Oncology: Basic Principles and Clinical Management of Cancer*, 2d ed., 1993; S. Cryz, *Immunotherapy and Intravenous Immunoglobulins*, 1991; T. J. Hamblin, *Immunotherapy and Disease*, 1990; P. Imbach (ed.), *Immunotherapy with Intravenous Immunoglobulins*, 1991; M. S. Mitchell, The advent of monoclonal antibodies, *Consultant*, 23:127–134, 1983.

Impact

A force which acts only during a short time interval but which is sufficiently large to cause an appreciable change in the momentum of the system on which it acts. The momentum change produced by the impulsive force is described by the momentum-impulse relation. For a discussion of this relation *see* IMPULSE (MECHANICS)

The concept of impulsive force is most useful when the time in which the force acts is so short that the system which it acts on does not move appreciably during this time. Under these conditions the momentum of the system is changed rapidly by a finite amount. The details of the way in which the force varies with time are unimportant, since only the impulse determines the momentum change.

Ordinarily the forces occurring in collisions are impulsive. In the processes in which impulsive forces occur, mechanical energy can be dissipated, and attempts to apply the conservation of mechanical energy may lead to incorrect results. *See* COLLISION (PHYSICS); CONSERVATION OF ENERGY.

A phenomenon known as impulsive loading occurs when materials are subjected to high-speed impacts or explosive charges. The study of failure of materials under impulsive loads is increasing in technological importance. *See* FORCE. Paul W. Schmidt

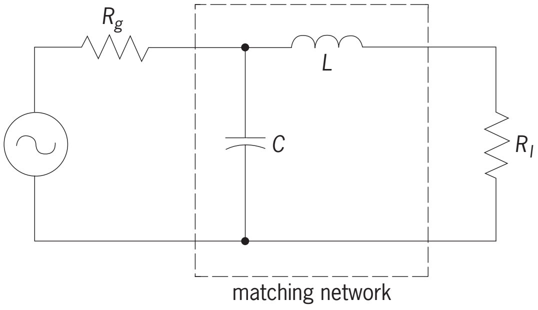
Bibliography. F. J. Bueche, *Introduction to Physics for Scientists and Engineers*, 4th ed., 1986; D. Halliday, R. Resnick, and K. Krane, *Fundamentals of Physics*, 6th ed., 2000; H. D. Young and R. A. Freedman,*Sears and Zemansky's University Physics*, 10th ed., 2000.

Impedance matching

The use of electric circuits and devices to establish the condition in which the impedance of a load is equal to the internal impedance of the source. This condition of impedance match provides for the maximum transfer of power from the source to the load. In a radio transmitter, for example, it is desired to deliver maximum power from the power amplifier to the antenna. In an audio amplifier, the requirement is to deliver maximum power to the loudspeaker. *See* ELECTRICAL IMPEDANCE.

The maximum power transfer theorem of electric network theory states that at any given frequency the maximum power is transferred from the source to the load when the load impedance is equal to the conjugate of the generator impedance. Thus, if the generator is a resistance, the load must be a resistance equal to the generator resistance for maximum power to be delivered from the generator to the load. When these conditions are satisfied, the power is delivered with 50% efficiency; that is, as much power is dissipated in the internal impedance of the generator as is delivered to the load.

Impedance matching network. In general, the load impedance will not be the proper value for maximum power transfer. A network composed of inductors and capacitors may be inserted between the load and the generator to present to the generator an impedance that is the conjugate of the generator impedance. Since the matching network is composed of elements which, in the ideal case of no resistance in the inductors and perfect capacitors, do not absorb power, all of the power delivered to the matching network is delivered to the load. An example of an L-section matching network is shown in the **illustration**. Matching networks of this type are used in radio-frequency circuits. The values of inductance and capacitance are chosen to satisfy the requirements of the maximum power transfer theorem. The power dissipated in the matching network is a small fraction of that delivered to the load, because the elements used are close approximations of ideal components.



L-section impedance matching network.

Transformers. The impedance measured at the terminals of one winding of an iron-cored transformer is approximately the value of the impedance connected across the other terminals multiplied by the square of the turns ratio. Thus, if the load and generator impedances are resistances, the turns ratio can be chosen to match the load resistance to the generator resistance for maximum power transfer. If the generator and load impedances contain reactances, the transformer cannot be used for matching because it cannot change the load impedance to the conjugate of the generator impedance (the L-section matching network can). The turns ratio can be chosen, however, to deliver maximum power under the given conditions, this maximum being less than the theoretical one.

Iron-cored transformers are used for impedance matching in the audio and supersonic frequency range. The power dissipated in the core increases with frequency because of hysteresis. Above the frequency range at which iron-cored transformers can be used, the air-core transformer or transformers with powdered-iron slugs can be used effectively. However, in these cases the turns-ratio-squared impedance-transforming property is no longer true. Since the transformer is usually part of a tuned circuit, other factors influence the design of the transformer.

The impedance-transforming property of an iron-cored transformer is not always used to give maximum power transfer. For example, in the design of power-amplifier stages in audio amplifiers, the impedance presented to the transistor affects distortion. A study of a given circuit can often show that at a given output power level, usually the maximum expected, there is a value for the load resistance which will minimize a harmonic component in the harmonic distortion, such as the second or third harmonic. The transformer turns ratio is selected to present this resistance to the transistor. *See* TRANSFORMER.

Emitter follower. In electronic circuitry a signal source of large internal impedance must often be connected to a low-impedance load. If the source were connected directly to the load, attenuation of the signal would result. To reduce this attenuation, an emitter follower is connected between the source and the load. The input impedance of the emitter follower is high, more nearly matching the large source impedance, and the output impedance is low, more nearly matching the low load impedance. If the object were the delivery of maximum power to the load, it might be possible to design the emitter follower to have an output resistance equal to the load resistance, assuming that the load is a resistance. (Special audio amplifiers have been designed to use emitter followers, rather than a transformer, to connect the loudspeaker to the power amplifier.) In many cases, maximum power transfer is not the goal; the emitter follower is introduced primarily to reduce to a minimum the attenuation of the signal.

There exist a number of applications where the emitter follower is not useful as an impedance matching circuit. For example, if a very low-impedance source must be matched to a high-impedance load, then a transistor is used in the common base configuration. *See* EMITTER FOLLOWER; VOLTAGE AMPLIFIER.

Christos C. Halkias

Bibliography. C. A. Schuler, *Electronics: Principles and Applications*, 5th ed., 1998; R. L. Thomas, *A Practical Introduction to Impedance Matching*, 1976; M. E. Van Valkenburg, *Network Analysis*, 3d ed., 1974.

Impsonite

A black, naturally occurring carbonaceous material having specific gravity 1.10–1.25 and fixed carbon 50–85%. Impsonite is infusible and is insoluble in carbon disulfide. A vein of impsonite 10 ft (3 m) wide occurs in La Flore County, Oklahoma, and there are other deposits in Arkansas, Nevada, Michigan, Peru, Argentina, Brazil, and Australia. The Peruvian impsonite may be derived from grahamite by weathering. The origin of impsonite is not well understood, but it appears to be derived from a fluid bitumen that polymerized after it filled the vein in which it is found. Impsonites are known to contain unusually high percentages of vanadium. *See* ASPHALT AND ASPHALTITE; WURTZILITE.

Irving A. Breger

Impulse (mechanics)

The integral of a force over an interval of time. For a force **F**, the impulse **J** over the interval from t_0 to t_1 can be written as Eq. (1). The impulse thus

$$\mathbf{J} = \int_{t_0}^{t_1} \mathbf{F}\, dt \qquad (1)$$

represents the product of the time interval and the average force acting during the interval. Impulse is a vector quantity with the units of momentum.

The momentum-impulse relation states that the change in momentum over a given time interval equals the impulse of the resultant force acting during that interval. This relation can be proved by integration of Newton's second law over the time interval from t_0 to t_1. Let **P** represent the momentum at time *t*, with $\mathbf{P}_0$ and $\mathbf{P}_1$ being the values of **P** at times t_0 and t_1, respectively. Then Eq. (2) holds. If, as is

$$\begin{aligned}\mathbf{J} &= \int_{t_0}^{t_1} \mathbf{F}\, dt = \int_{t_0}^{t_1} (d\mathbf{P}/dt)\, dt \\ &= \int_{\mathbf{P}_0}^{\mathbf{P}_1} d\mathbf{P} = \mathbf{P}_1 - \mathbf{P}_0 \end{aligned} \qquad (2)$$

ordinarily true, the mass *m* is constant, the momentum change can be expressed in terms of the velocities $\mathbf{v}_1$ and $\mathbf{v}_0$ at times t_1 and t_0, respectively, giving Eq. (3).

$$\mathbf{J} = m(\mathbf{v}_1 - \mathbf{v}_0) \qquad (3)$$

The concept of impulse is ordinarily most useful when the forces are large but act only for a short period. In most of these cases it is necessary to know only the momentum change, which is determined by the impulse. The relation between momentum and impulse thus has the advantage of eliminating the need for a detailed knowledge of how the forces, which can be very complicated, change with time. Forces which occur during collisions are of this type. *See* COLLISION (PHYSICS).

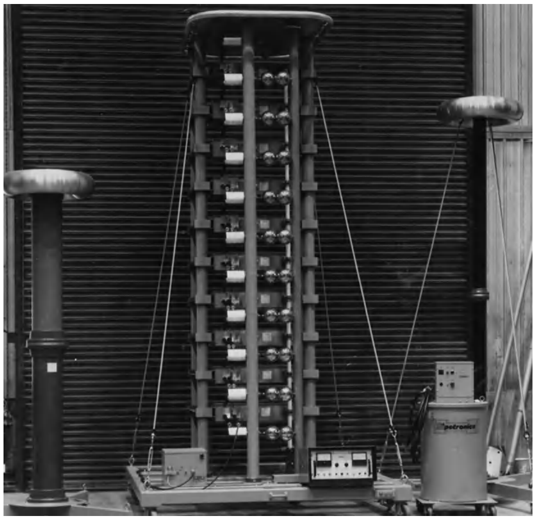
A further simplification of this type of system is obtained if the time interval is short enough for the system to be considered essentially stationary during the time of action of the force, so that the momentum change occurs almost instantaneously. Except for the changed momentum, the motion of the system is treated as it would have been if no impulse had occurred. The effect of the impulse can thus be considered to provide a set of initial conditions for the motion. In a ballistic galvanometer, for example, there is an electric current for only a short time, during which the galvanometer coil, originally at rest, is given an impulse by the force associated with the current. The only effect of the current on the later motion of the coil, which can be considered stationary during the time there is a current, is to provide an initial velocity for the subsequent motion. *See* GALVANOMETER; IMPACT; MOMENTUM.

Paul W. Schmidt

Bibliography. F. Bueche, *Introduction to Physics for Scientists and Engineers*, 4th ed., 1986; D. Halliday, R. Resnick, and K. Krane, *Fundamentals of Physics*, 6th ed., 2000.

Impulse generator

An electrical apparatus which produces very short high-voltage or high-current surges. Such devices can be classified into two types: impulse voltage generators and impulse current generators. High impulse voltages are used to test the strength of electric power equipment against lightning and switching surges. Also, steep-front impulse voltages are sometimes used in nuclear physics experiments. High impulse currents are needed not only for tests on equipment such as lightning arresters and fuses but also for many other technical applications such as lasers, thermonuclear fusion, and plasma devices. *See* FUSE (ELECTRICITY); LASER; LIGHTNING AND SURGE PROTECTION; NUCLEAR FUSION; PARTICLE ACCELERATOR; PLASMA (PHYSICS).



A 1-MV impulse voltage generator. (*Hipotronics, Inc.*)

Impulse voltage generator. An impulse voltage generator (sometimes called a Marx generator, after E. Marx who first proposed it in 1923) consists of capacitors, resistors, and spark gaps. The capacitors are first charged in parallel through charging resistors by a high-voltage, direct-current source and then connected in series and discharged through a test object by a simultaneous spark-over of the spark gaps. The output voltage of such a generator is determined by the number of capacitors and the charging voltage. It is also determined by the efficiency of the generator. The waveform of the impulse voltage can be controlled by suitably adjusting the circuit components of such a generator.

The fundamental parameters of an impulse voltage generator are its rated voltage, energy, waveform T_f/T_t, and polarity. The wavefront time T_f, in microseconds, represents the steepness of the front of the impulse wave. The wave-tail time T_t, defined as the time, in microseconds, required for the tail of the wave to decline to half its peak value, represents the speed of attenuation of the impulse wave. The value of T_f/T_t for a standard lightning impulse wave is 1.2/50, and for a switching impulse wave it is 250/2500. The highest rated voltage of an impulse voltage generator is 10 MV.

The 1-MV impulse voltage generator shown in the **illustration** consists of 10 stages. Each stage has one capacitor. The maximum charging voltage of these capacitors is 100 kV.

Impulse current generator. The impulse current generator comprises many capacitors that are also charged in parallel by a high-voltage, low-current, direct-current source, but it is discharged in parallel through resistances, inductances, and a test object by a spark gap. It provides very high current. Its parameters are rated amperage, waveform T_f/T_t, and polarity. For standard impulse current waves, the values of T_f/T_t, in microseconds, are 4/10 and 8/20. The waveform can be controlled by suitably adjusting the value of the circuit resistances and inductances. The highest rated amperage of an impulse current generator is above several hundred kiloamperes.

Changchang Wang; T. C. Cheng

Bibliography. T. J. Gallagher and A. J. Pearmain, *High Voltage Measurement, Testing and Design*, 1983; H. M. Ryan, *High Voltage Engineering and Testing*, 1994.

Impulse turbine

A turbine in which fluid is deflected without a pressure drop in the blade passages. A turbine is a power-producing machine fitted with shaft-mounted wheels. Turbine blades, attached to the wheels' periphery, are driven by the through-flow of water, steam, or gas (**Fig. 1**). The rotary motion of the wheel is maintained by forces imparted to the blades by the impingement against them of high-speed fluid streams. Before the stream of fluid reaches the moving turbine blades, it is accelerated in stationary

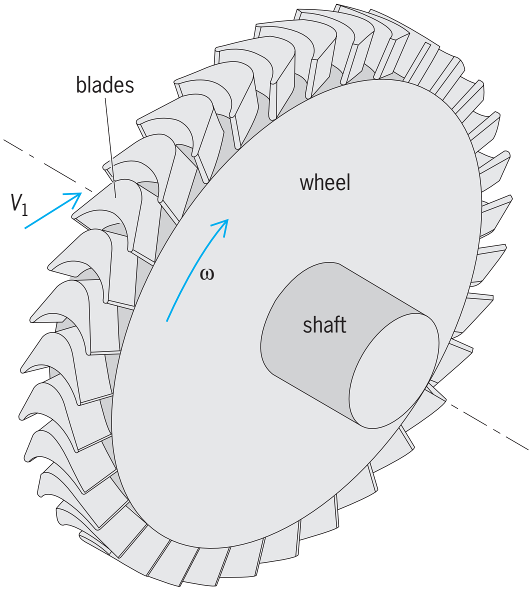


Fig. 1. Impulse turbine wheel. Symbols are explained in text.

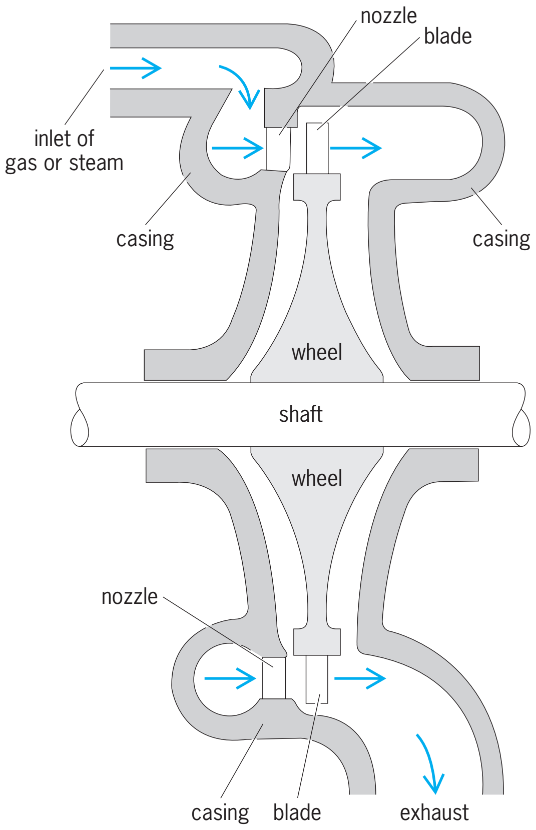


Fig. 2. Longitudinal cross section of a steam or gas impulse turbine.

passages called nozzles (**Fig. 2**). The nozzles are shaped to convert mechanical or thermal energy of the fluid into kinetic energy; that is, the nozzles increase the fluid's velocity while decreasing its pressure and temperature. Upon leaving the nozzles the high-speed fluid strikes the moving blades, and a force is imparted to the blades as the fluid is deflected by them. If the fluid's deflection in the blade passage is accompanied by a pressure drop and a relative velocity rise, the turbine is called a reaction turbine; if the fluid is deflected without a pressure drop in the blade passages, it is called an impulse turbine. *See* NOZZLE; REACTION TURBINE.

The principle of the impulse turbine derives from Newton's laws of motion as applied to the particles of fluid deflected by the blades. From application of the laws of motion to the impulse turbine's blades, it follows that the force of the fluid on the blades is equal and opposite to that of the blades on the fluid. Furthermore, Newton's principle states that the force on the fluid equals the time rate of change of its linear momentum, where linear momentum is defined as the fluid's mass m times its velocity V relative to the moving blade. This principle can also be stated in another way, namely, that the impulse given to the fluid is equal to its change of linear momentum, where impulse is defined as the product of average blade force F by the time t that the fluid resides in the blade passage. The mass m of the parcel will be the mass of fluid entering the blade passage during the time t.

If the fluid approaches the passage between the blades with a relative velocity V_1 and is moving, relative to the blades, in a direction which makes an angle θ with a tangent line to the circumference of the wheel (**Fig. 3**), then the tangential component of the momentum of an entering parcel of fluid of mass m is simply $mV_1 \cos\theta$. Similarly, the tangential component of momentum of the same fluid mass is $-mV_2 \cos\theta$ when it leaves the blade passages. Typically, impulse turbine blades are symmetrical in shape (Figs. 2 and 3), and there is no change in the magnitude of the relative velocity in the blade passage (if frictional effects are neglected); thus, $V_1 = V_2$, and the tangential component of the impulse, Ft, created by the deflection of the fluid in the blade passage is $2mV_1 \cos\theta$.

When an impulse turbine is driven with water, it is called a Pelton wheel (**Fig. 4**). In this case the impulse blade is shaped differently from that used in steam and gas turbines (Figs. 1 and 3), and the water stream impinges tangentially on the blade's center. Nevertheless, Pelton wheel blades are also

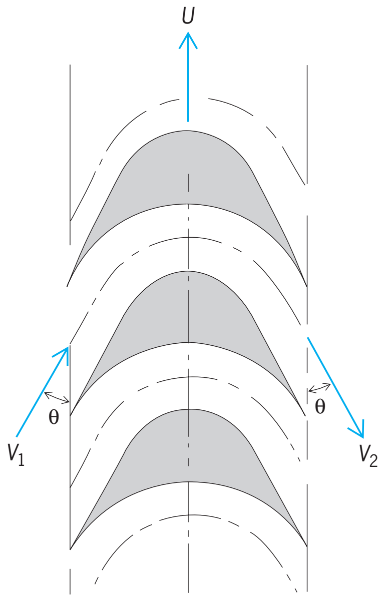


Fig. 3. Cross section through blades mounted at the periphery of the turbine wheel, showing impulse blading. Symbols are explained in text.

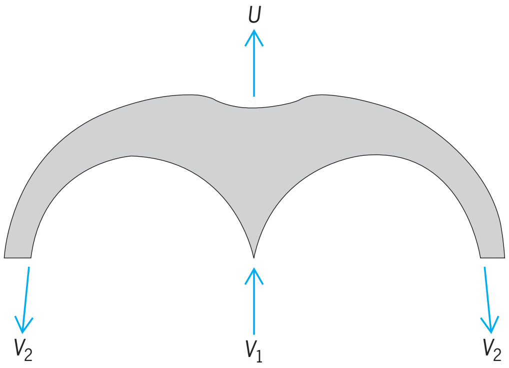


Fig. 4. Hydraulic impulse turbine blade of a Pelton wheel. Symbols are explained in text.

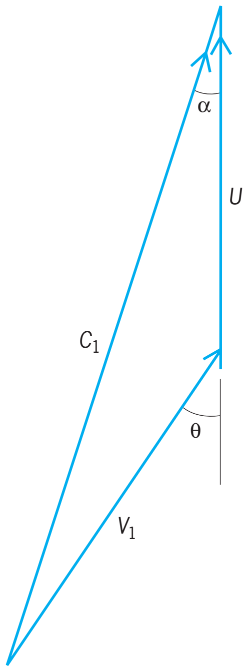


Fig. 5. Velocity diagram, showing relation of blade velocity U, absolute velocity C_1 of the stream entering the blade passages, and velocity V_1 of the entering stream relative to the blades.

symmetrical, and the equality, $V_1 = V_2$, applies to them as well.

If all the blades of the turbine wheel are considered together, then the the total blade force F created on all the blades by the total flow is given by $2mV_1 \cos\theta/t$; m/t is the mass rate of flow of fluid through the turbine. Since the velocity V_1 of the stream entering the blade passages is a velocity relative to the moving blade, it is related to the blade velocity U by the equation below. Here, the entering

$$V_1 \cos\theta + U = C_1 \cos\alpha$$

stream has an absolute velocity whose magnitude is C_1 and whose direction makes an angle α with a tangent to the circumference of the blades (**Fig. 5**). Since the stream entering the blade passages has just emerged from the nozzle, C_1 is also the nozzle exit velocity. Thus, $V_1 \cos\theta$ will be zero when the tangential component of the absolute velocity of the stream, $C_1 \cos\alpha$, equals the blade velocity U. The blade velocity is related to the shaft's angular velocity ω and the blade's average radial distance R from the shaft's axis of rotation; thus, the blade velocity is given by $U = \omega R$. When a turbine reaches a certain maximum shaft speed ω_m at which $\omega_m R = C_1 \cos\alpha$, no blade force is created, and the turbine cannot produce any work at all. Since the power P produced by the turbine is the blade force times the blade velocity, that is, $P = FU$, the turbine power is zero at both zero blade velocity ($\omega = 0$) and at zero blade force, that is, where the shaft speed $\omega = \omega_m$. Maximum power and maximum efficiency are realized between these two extremes of rotational speed. In fact, ideal turbine speed occurs when the blade velocity is approximately half the nozzle exit velocity C_1. *See* GAS TURBINE; HYDRAULIC TURBINE; PRIME MOVER; STEAM TURBINE; TURBINE.

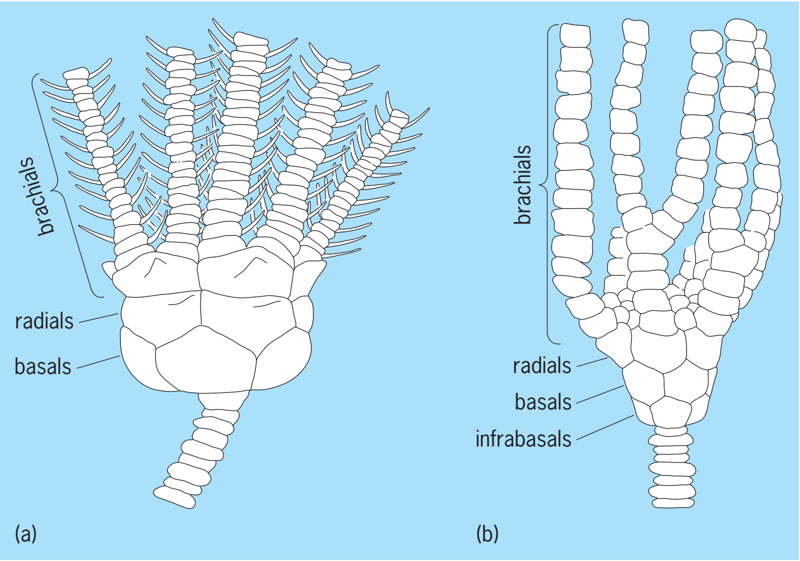
Earl Logan, Jr.

Bibliography. H. Cohen, G. F. C. Rogers, and H. I. H. Saravanamuttoo, *Gas Turbine Theory*, 4th ed., 1996; E. Logan (ed.), *Handbook of Turbomachinery*, 1994; F. M. White, *Fluid Mechanics*, 4th ed., 1999; D. G. Wilson and T. Korakianitis, *The Design of High-Efficiency Turbomachinery and Gas Turbines*, 2d ed., 1998.

Inadunata

One of three Paleozoic subclasses of the Crinoidea consisting of over 350 genera ranging from the Ordovician to the Permian. Its members generally lack fixed brachials such that the arms are free above the level of the radials. The calyx may be monocyclic, consisting of a circlet of basal and a circlet of radial plates (**illus.** *a*), or dicyclic, having an additional circlet of infrabasal plates (illus. *b*). The arms, constructed of serially arranged brachials, may be simple or branched. Each arm plate may be bare or give rise to a slender, unbifurcated branchlet called a pinnule. Articulating surfaces of adjacent brachials may be smooth, forming a tight, inflexible suture, or each may possess a central ridge. In the latter case, abutting ridges of adjacent brachials form a mechanical fulcrum allowing a substantial degree of movement and flexibility. The anus may lie flush with the tegmen or be elevated on a tube of variable length and diameter.

The generic diversity of the Inadunata reached its peak during the Late Paleozoic, when it became the dominant crinoid group; all post-Paleozoic crinoids, members of the subclass Articulata, descended from one of the clades of the Inadunata. Cladistic analyses of the Inadunata reveal a lack of shared derived characters (synapomorphies), implying that



Inadunates. (*a*) Monocyclic. (*b*) Dicyclic.

the taxon represents a polyphyletic grouping of distantly related crinoid clades. *See* CRINOIDEA; ECHINODERMATA.

Tomasz K. Baumiller

Bibliography. W. I. Ausich, Phylogeny of Arenig and Caradoc crinoids (phylum Echinodermata) and supergeneric classification of the Crinoidea, *Univ. Kans. Paleontol. Contrib. New Ser.*, 9:1–36, 1998; H. Hess et al. (eds.), *Fossil Crinoids*, Cambridge University Press, Cambridge, 1999; R. C. Moore and C. Teichert (eds.), *Echinodermata 2: Crinoidea*, pt. T of *Treatise on Invertebrate Paleontology*, 1978; C. R. C. Paul and A. B. Smith (eds.), *Echinoderm Phylogeny and Evolutionary Biology*, Clarendon Press, London, 1988; M. Simms and G. D. Sevastopulo, The origin of articulate crinoids, *Palaeontology*, 36:91–109, 1993.

Inarticulata

Formerly, one of the two classes of the phylum Brachiopoda. The phylum comprises solitary, exclusively marine, coelomate [having a true coelom or mesodermally lined body cavity], lophophorate [having a lophophore, a food-gathering ciliated organ], bivalved animals, with the two valves symmetrical about a median longitudinal plane. Now the former Inarticulata comprises two of the three subphyla currently recognized, Linguliformea and Craniiformea, together with a few taxa considered to be basal rhynchonelliform brachiopods. These two subphyla encompass approximately 6% of the generic diversity of the Recent (Holocene, from 10,000 years ago to the present day) and fossil brachiopods. The name Inarticulata refers to the lack of tooth-and-socket articulatory structures on the two valves of these brachiopods. The name was abandoned because of recent changes made in the classification thought to reflect a better understanding of phylogenetic relationships among several groups of Cambrian brachiopods. Because some linguliform and craniiform species have primitive forms of articulation that may or may not be homologous with articulated brachiopods, and because some early rhynchonelliforms lack articulation, the term Inarticulata no longer accurately characterizes this group of brachiopods. However, molecular and morphological evidence suggests that Linguliformea and Craniiformea are each likely to be monophyletic (evolved from a single interbreeding population), are sister groups to one another, and may form a clade, together with Phoronida at the base, that is the sister group to Rhynchonelliformea. *See* BRACHIOPODA; PHORONIDA; RHYNCHONELLIFORMEA.

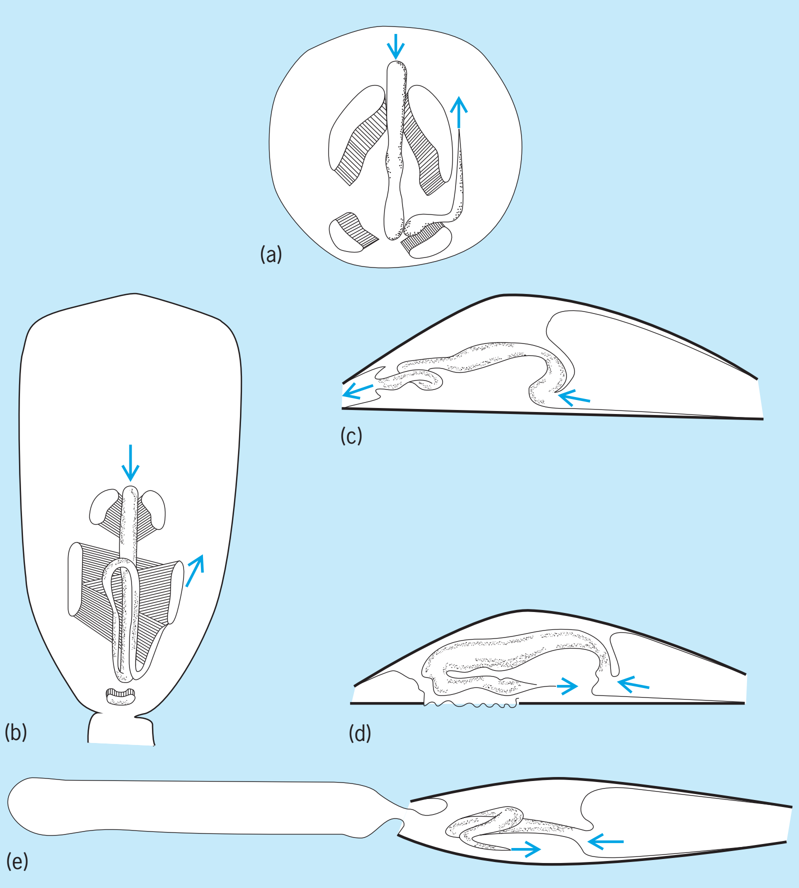


Fig. 1. Position of the gut in relation to the major muscles between the valves; position of mouth and anus and direction of movement of material indicated by arrows. (*a*) *Discinisca*, dorsal view, as if dorsal valve is transparent. (*b*) *Lingula*, dorsal view, as if dorsal valve is transparent. (*c–e*) Comparisons between lateral views of three of the four main types of brachiopod gut arrangement: (c) *Novocrania*; (*d*) *Discinisca*; (*e*) *Lingula*. (*After R. L. Kaesler, ed., Treatise on Invertebrate Paleontology, Part H, Revised, vol. 1, Geological Society of America and University of Kansas Press, 1997; adapted from Nielsen, 1991*)

Classification is as follows:

- Phylum Brachiopoda
  - Subphylum Linguliformea
    - Class Lingulata
      - Order: Lingulida
        - Acrotretida
        - Siphonotretida
    - Class Paterinata
      - Order: Paterinida
  - Subphylum Craniiformea
    - Class Craniata
      - Order: Craniida
        - Craniopsida
        - Trimerellida
  - Subphylum Rhynchonelliformea

Living linguliforms and craniiforms share a number of morphological features that can distinguish them from rhynchonelliforms. These features include a mantle (ectodermal epithelium) that does not reverse (turn "inside out") in development; discrete, not continuous mantle epithelium at the posterior of the animal; valve articulation that is largely but not exclusively absent; lack of mineralized lophophore supports; and presence of an anal opening (these last two features appear to be primitive characters shared with a common ancestor).

Linguliformea. Linguliformea may (artificially) unite two distinct groups, lingulates and paterinates, and is further characterized by the presence of a pedicle (a tough but flexible cuticle-covered structure at or near the posterior of the animal) that has a coelomic core that develops from the ventral mantle, not as a larval rudiment as in rhynchonelliforms. An organophosphatic exoskeleton (of two halves) with a stratiform structure with vertical canals is present in all linguliforms.

Lingulata. The lingulates are represented in the modern marine fauna by lingulids and discinids; several genera of each are cosmopolitan (having worldwide range) in their biogeographical distribution. *Lingula* and *Glottidia* have a free-living lifestyle, unusual for brachiopods, and burrow in soft substrata, apparently by contracting dermal muscles and utilizing changes in hydrostatic pressure within the coelom. *Discinisca* lives attached to hard substrata by a pedicle, and possesses a bimineralic (silica and organophosphate) shell, with minute siliceous tablets formed as an early precursor to the organophosphatic shell. Lingulates have planktotrophic (feeding) larvae, which are essentially swimming juveniles with shells, spending months in the water column, and have gonads located in the body cavity, rather than the mantle canals as in rhynchonelliforms. Lingulids can tolerate reduced salinity and can be found today and in the fossil record in marginal marine-to-brackish, shallow-water environments. Lingulate valves can vary from subcircular to elongate oval to subrectangular in outline, and from gently to steeply conical (**Fig. 1***a*, *b*, *d*, *e*; **Fig. 2***b*, *d*, *e*) in lateral view. Their muscle systems (**Fig. 3**) involve several pairs of oblique muscles, in addition to adductor (closing) muscles; they typically do not have diductor (opening) muscles, as rhynchonelliforms do. Acrotretides and siphonotretides are two extinct groups of early Paleozoic linguliforms. Acrotretides have conical ventral valves with large pseudo-interareas (somewhat flattened, posterior shell sectors) [Fig. 2*e*] and an apical pedicle foramen (opening). Siphonotretides have hollow spines on their valve exterior and a large apical pedicle foramen. *See* ACROTRETIDA; LINGULIDA.

Paterinata. Paterinates, the first brachiopods to appear in the fossil record [in the Lower Cambrian (Tommotian)], are known only as fossils from the Cambrian and Ordovician periods. They are

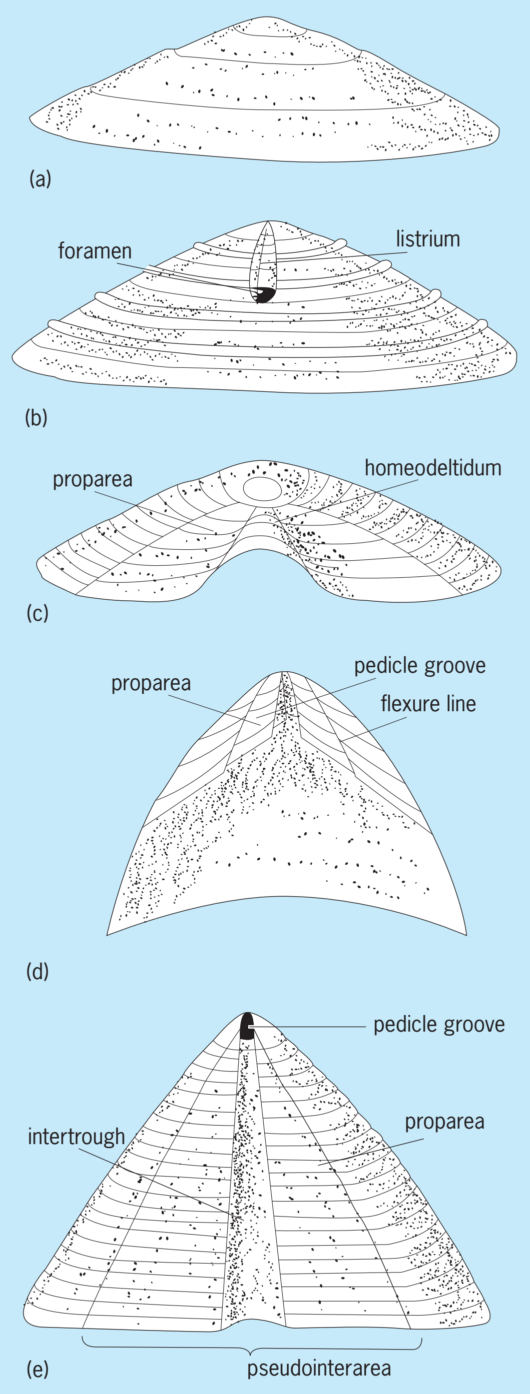


Fig. 2. Modifications of the posterior sector of the ventral valve of some linguliform and craniiform brachiopods. (*a*) Craniiform *Pseudocrania*. (*b*) Discinoid *Orbiculoidea*. (*c*) Paterinate *Paterina*. (*d*) Linguloid *Lingulella*. (*e*) Acrotretide *Prototreta*. (*After R. L. Kaesler, ed., Treatise on Invertebrate Paleontology, Part H, revised, vol. 1, Geological Society of America and University of Kansas Press, 1997*)

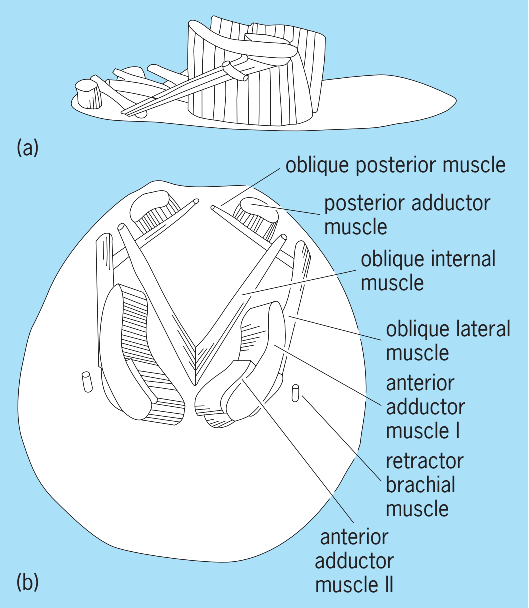


Fig. 3. Muscle system of the discinoid linguliform *Discinisca*. (*a*) Lateral view. (*b*) Dorsal view. (*After R. C. Moore, ed., Treatise on Invertebrate Paleontology, Part H, Geological Society of America and University of Kansas Press, 1965*)

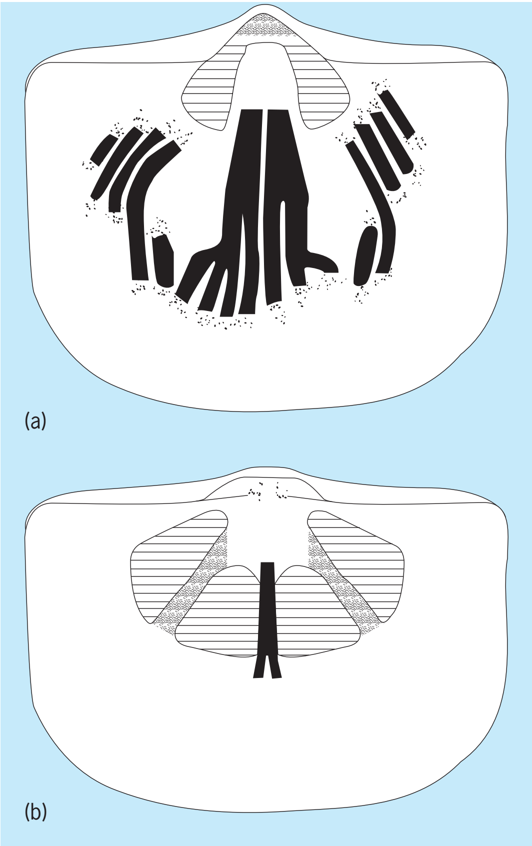


Fig. 4. Interpretation of musculature (muscle scars are hatched) and mantle canal systems (black) of the paterinate brachiopod *Askepasma*. (*a*) Ventral valve interior. (*b*) Dorsal valve interior. (*After R. L. Kaesler, ed., Treatise on Invertebrate Paleontology, Part H, revised, vol. 2, Geological Society of America and University of Kansas Press, 2000*)

morphologically more similar to early rhynchonelliforms than to other linguliforms, differing primarily in the composition of the shell. Their valves have a straight (strophic) posterior margin (Fig. 2*c*; **Fig. 4**) and lack articulation, but appear to have diductor (valve opening) muscle scars (impressions or elevations on valves representing the final sites of attachment of muscles); their valves may well have rotated relative to one another about a hinge axis in life. Paterinates may prove to be more closely related to rhynchonelliforms if organophosphatic mineralogy has evolved more than once in brachiopods. This is certainly a possibility given our incomplete understanding of metazoan mineralization early in the Cambrian. *See* PATERINIDA.

Craniiformea. Craniiformea is characterized by the lack of a pedicle at any stage in development. Recent adults are cemented to hard substrates by their ventral valves. Both valves are typically subcircular in outline and are very slightly conical (Fig. 1*c*, 2*a*) to planar in lateral view. They have lecithotrophic (nonfeeding) rather than planktotrophic larvae, and endopunctate calcareous, not canaliculate phosphatic, valves with a distinctive laminar (tabular) shell structure. All living craniates lack articulation, but many have strophic posterior valve edges. *Novocrania* appears to have a cosmopolitan distribution. Craniopsides, extinct Paleozoic craniiforms, are cemented or free-lying, with impunctate (lacking punctate) valves. Trimerellides, extinct free-lying early Paleozoic craniiforms, have rudimentary articulatory structures in the valves, muscle scars on raised platforms, and unusually large and thick shells, possibly of aragonitic rather than calcitic composition.

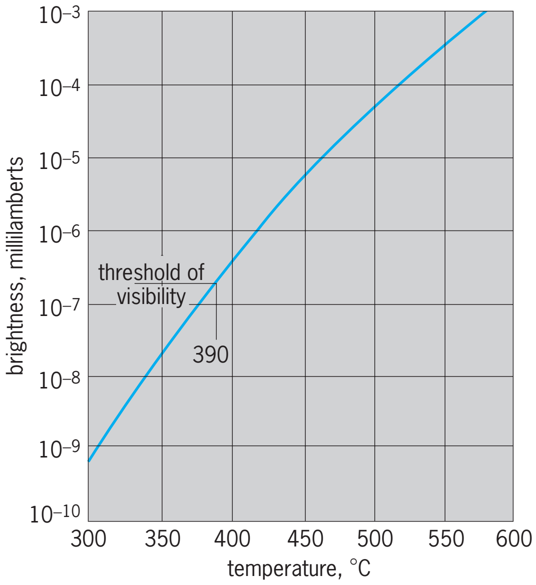
Sandra J. Carlson

Bibliography. S. J. Carlson, Recent research on brachiopod evolution, Part H: Brachiopoda, revised, vol. 6: Supplement, in R. L. Kaesler (ed.), *Treatise on Invertebrate Paleontology*, Geological Society of America and University of Kansas, 2007; B. L. Cohen and A. Weydmann, Molecular evidence that phoronids are a subtaxon of brachiopods (Brachiopoda: Phoronata) and that genetic divergence of metazoan phyla began long before the Early Cambrian, *Organisms, Diversity & Evolution*, 5(4):253–273, 2005; B. L. Cohen, A. Gawthrop, and T. Cavalier-Smith, Molecular phylogeny of brachiopods and phoronids based on nuclear-encoded small subunit ribosomal RNA gene sequences, *Phil. Trans. Roy. Soc. London*, Series B, 353(1378):2039–2061, 1998; A. Williams et al., Part H: Brachiopoda, revised, vol. 2: Linguliformea, Craniiformea, and Rhynchonelliformea (part), in R. L. Kaesler (ed.), *Treatise on Invertebrate Paleontology*, Geological Society of America and University of Kansas, 2000; A. Williams and S. J. Carlson, Affinities of brachiopods and trends in their evolution, Part H: Brachiopoda, Revised, vol. 6: Supplement, in R. L. Kaesler (ed.), *Treatise on Invertebrate Paleontology*, Geological Society of America and University of Kansas, 2007; A. Williams and A. J. Rowell, Classification, Part H: Brachiopoda, vol. 1, in R. C. Moore (ed.), *Treatise on Invertebrate Paleontology*, Geological Society of America and University of Kansas, H214–H236, 1965.

Incandescence

The emission of visible radiation by a hot body. A theoretically perfect radiator, called a blackbody, will emit radiant energy according to Planck's radiation law at any temperature. Prediction of the visual brightness requires additional consideration of the sensitivity of the eye, and the radiation will be visible only for temperatures of the blackbody which are above some minimum. The relation between brightness and temperature is plotted in the **illustration**. As shown, the minimum temperature for incandescence for the dark-adapted eye is about 730°F (390°C). Under these ideal observing conditions, the incandescence appears as a colorless glow. The dull red light commonly associated with incandescence of objects in a lighted room requires a temperature of about 930°F (500°C). *See* BLACKBODY; HEAT RADIATION; VISION.

Not all sources of light are incandescent. A cold gas under electrical excitation may emit light, as in the so-called neon tube or the low-pressure mercury-vapor lamp. Ultraviolet light from mercury vapor may excite visible light from a cold solid, as in the fluorescent lamp. Luminescence is the term used to refer to the emission of light due to causes other



Graph showing the relation between the brightness of a blackbody and temperature. °F = (°C × 1.8) + 32.

Approximate color temperatures of common light sources

Source	Color temperature, °F (K)
Candle	3005 (1925)
Kerosine lamp	3140 (2000)
Common tungsten-filament 100-watt electric light bulb	4600 (2800)
Carbon arc	6700 (4000)
Sun	10,000 (5800)

than high temperature, and includes thermoluminescence, in which emission of previously trapped energy occurs on moderate heating. *See* FLUORESCENT LAMP; MERCURY-VAPOR LAMP; THERMOLUMINESCENCE.

Flames are made luminous by incandescent particles of carbon. Gas flames can be made to produce intense light by the use of a gas mantle of thoria containing a small amount of ceria. This mantle is a good emitter of visible light, but a poor emitter of infrared radiation. As less heat is lost in the long waves, the mantle operates at a higher temperature than a blackbody would and, hence, produces more intense visible light.

A useful criterion of an incandescent source is its color temperature, the temperature at which a blackbody has the same color, although not necessarily the same brightness. The color temperature of common light sources depends upon operating conditions. Approximate values are given in the accompanying **table**. *See* INCANDESCENT LAMP.

H. W. Russell; George R. Harrison

Bibliography. F. A. Jenkins and H. E. White, *Fundamentals of Optics*, 4th ed., 1976; J. R. Meyer-Arendt, *Introduction to Classical and Modern Optics*, 4th ed., 1994; Optical Society of America, *Handbook of Optics*, 2d ed., 1994.

Incandescent lamp

A lamp that creates radiant energy when its metallic filament is heated by an electric current. The filament is designed to produce radiant energy in the visible portion of the electromagnetic spectrum (light). The filament is of a special material that is supported in an envelope (bulb) that has been evacuated or filled with an inert gas such as argon, nitrogen, or krypton. In addition to light, the heated filament emits infrared and ultraviolet energy. When either of these radiations is accentuated, the lamp may be used as a source of that energy but with a reduction in luminous efficacy. The luminous efficacy (formerly called light-source efficiency) is the ratio of the units of light produced (lumens) to the power (watts) required, and is expressed in units of lumens per watt (lm/W). *See* LIGHT; LUMINOUS EFFICACY.

Advantages and disadvantages. The advantages of incandescent lamps include the following: (1) Lamps are low in cost compared to other illumination devices. (2) Lamps are simple to use and to connect into the electrical system. (3) Lamps need no auxiliary, such as a ballast. (4) The light source may be kept very compact, making the light easily controlled by reflectors and lenses. (5) Many shapes and sizes are available for special uses. (6) Color rendition of skin colors is pleasing to the observer. (7) Lamps can be designed to operate at low voltage, that is, 1.5 V, and are therefore readily used with battery sources of power (flashlights and autos). (8) Inexpensive dimmer control of light output is possible.

The disadvantages include the following: (1) Efficacy is low compared to fluorescent or high-intensity-discharge lamps. (2) Lamps are very inefficient when colored light at the blue end of the spectrum is required. (3) Lamps have short life compared to fluorescent and high-intensity-discharge lamps. (4) A large amount of heat is produced.

Lamp construction. The important parts of an incandescent lamp are the bulb (envelope), the filament, and the base (**Fig. 1**).

The bulb may be clear, colored, inside-frosted, or coated with diffusing or reflecting material. Most lamps have soft-glass bulbs; hard glass is used when

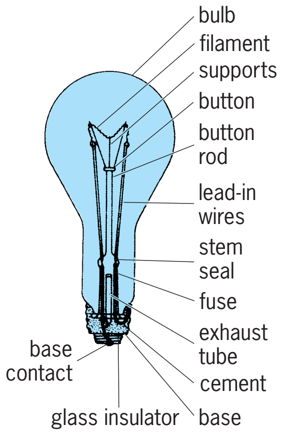


Fig. 1. Parts of an incandescent lamp. (*General Electric Co.*)

the lamp will be subjected to sudden and severe temperature changes. Lamps have a variety of bulb shapes, base types, and filament structures (**Fig. 2**). These vary according to the type of service planned, the need for easy replacement, and other environmental and service conditions.

The efficient design of an incandescent lamp centers about obtaining a high temperature at the filament without the loss of heat or disintegration of the filament. The early selection of carbon, which has the highest melting point of any element (3872 K or 6510°F) was a natural one. (It is standard practice to express filament temperatures in kelvins. The Kelvin temperature scale has degrees identical in size to the Celsius degree; however, the scale, instead of starting at the melting point of ice, as does the Celsius scale, starts at absolute zero, which occurs at −459.67°F or −273.15°C.) Carbon evaporates from its solid phase (sublimates) below this temperature, so carbon filaments must be operated at relatively low temperatures to obtain reasonable life. With the advent of ductile tungsten (1911) a nearly perfect filament material was discovered. Ductile tungsten has a tensile strength four times that of steel, its melting point is high (3655 K or 6120°F), and it has relatively low evaporation. Hot tungsten is an efficient light radiator; it has a continuous spectrum closely following that of a blackbody radiator with a relatively high portion of the radiation in the visible spectrum. Because of its strength, ductility, and workability, it may be formed into coils, these coils again recoiled (for coil-coil filaments), and these again recoiled for cathodes in fluorescent lamps. If tungsten could be held at its melting point, 52 lm/W would be radiated. Because of physical limitations, however, 22 lm/W is the highest practical radiation for general-service lamps; some special lamps can reach as high as 35.8 lm/W. In a given incandescent lamp, a higher filament temperature will result in an increased efficacy and color temperature (whiter light) and a reduction in lamp life.

Vaporization of the filament is reduced as much as possible. A small amount remains, however, and causes blackening of the bulb. The evaporated tungsten particles are carried to the upper part of the bulb by convection currents. In a base-down position the blackening reduces the output a few percent. To reduce blackening, the inner atmosphere of the lamp is maintained as clean as possible by use of a getter, which combines chemically with the tungsten particles. In some lamps a grid is placed to attract and hold the evaporated tungsten particles.

The electric power supplied to the filament of a general service lamp is converted to radiant energy; however, only 4–6% of the energy is light. The remaining energy is heat radiation, except for a very minor part which is in the form of ultraviolet radiation. This better than 90% conversion to heat results in portions of the lamp operating at rather high temperatures. The 100-W general service lamp may have glass temperatures of almost 450°F (230°C) operated base-down and base temperatures over 225°F (108°C) operated base-up. Placing the lamp in some types of luminaires (light fixtures) or other devices that tend to hold the heat in may produce lamp temperatures well above these. Excessive heat may cause the cement holding the base to the glass to fail. This failure can be prevented by selecting lamps with skirted or bipost bases designed for use at high ambient temperatures (Fig. 2). Lamp filaments that require large electric currents, such as high-wattage 120-V lamps or moderate-wattage 6-V lamps, are designed with bipost or other bases that can be solidly connected to the electric circuit.

Slide and movie projector lamps are examples of lamps required to operate at high temperatures. A common failure of such lamps results from improper adjustment of the projector optics. The filament image is focused on the bulb wall instead of the film and the glass softens and blisters. Another common failure results when a lamp is exposed to a spray or mist of cold water, causing thermal cracks in the glass. To protect the electric circuit from some lamp failures, a fuse is built into the lead wire that connects the base to the filament.

Lamp ratings. Most lamps are rated in watts at a specified voltage. The most common voltage is 120 V for lamps to be used for general lighting service. General-service lamps are also available with voltage ratings of 125, 130, 230, 250, and 277. Lamps for special lighting service where the voltage may not be relatively constant may have ratings such as 115–125 V. Typical of such lamps are the rough-service lamps for use on extension cords where supplementary lighting is needed. Lamps for various special uses have voltage ratings which range from 1.5 V for flashlight lamps, 6 V for projector lamps, 12 V for automotive lamps, to 300 V for mines and special industrial usage.

Decorative lamps used in strings for Christmas trees and similar uses are available in two forms. The first uses 7–15 W 120-V lamps connected so the lamps operate in parallel from a 120-V source of power. The second form uses 35–50 miniature lamps connected in series to a 120-V source. The lamps are rated at 2.5–3.5 V and normally have currents in the range 0.04–0.06 A. A device which shorts the lamp when a filament burns out keeps the entire string from going out when one or two lamps burn out, but if three or more lamps are burned out the entire string is extinguished. The power for an entire string of these miniature lamps is only 5–7 W.

Lamp characteristics. Among the more important lamp characteristics are life, lumen (light) output, efficacy or lumens per watt of input power, color, color temperature, lumen depreciation, and the effect of variations in the applied voltage on the above characteristics of lamps.

Life and efficacy. The life of an incandescent lamp may range from a low 2–6 h for a photoflood lamp, where life has been sacrificed to obtain high efficacy and high color temperature, to small long-life lamps designed for exit fixtures. The latter lamps have low efficacy but a very long 10,000–100,000-h life to assure continuity of service. General-service

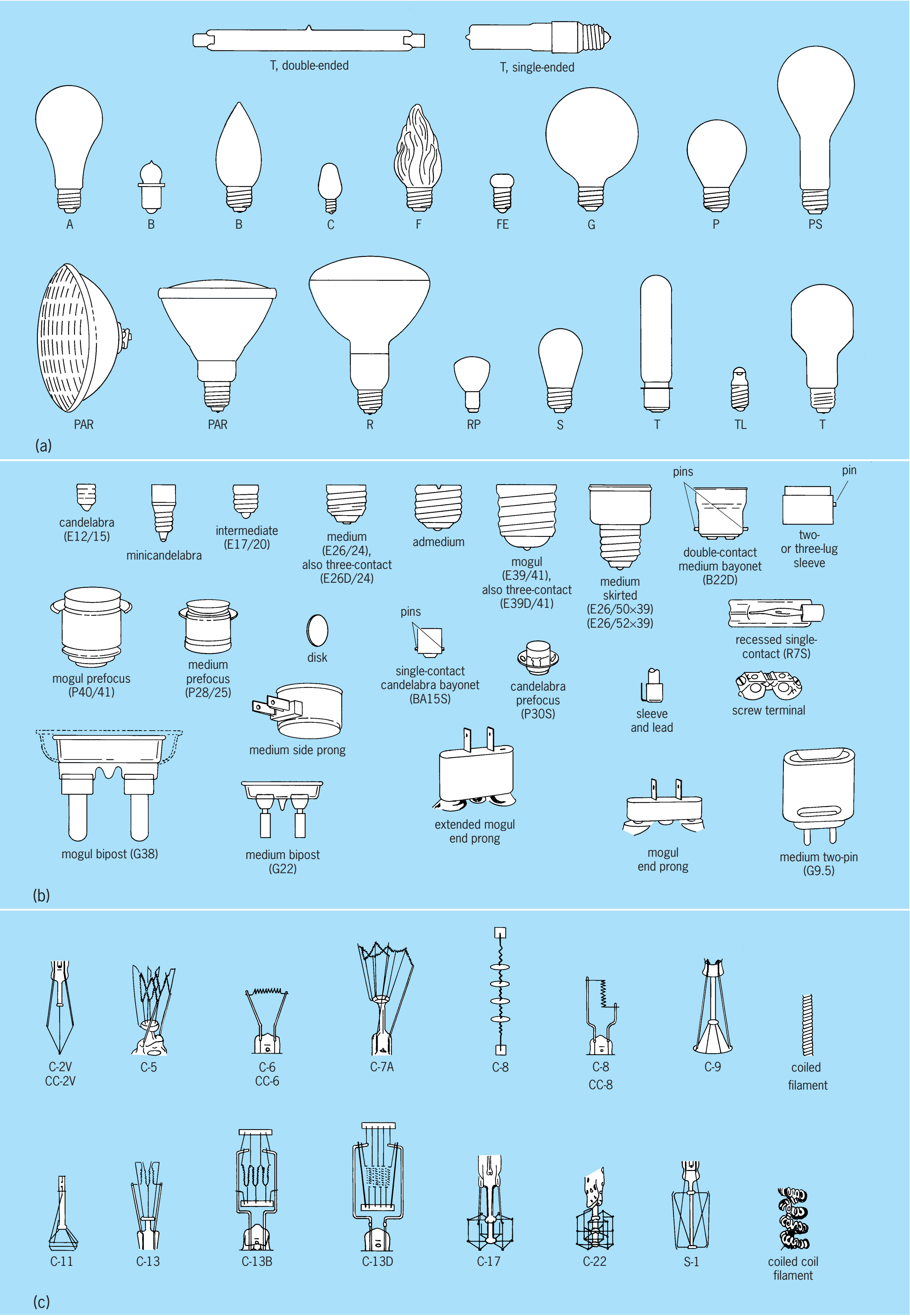


Fig. 2. Variations of incandescent lamps. Drawings are not to scale. (*a*) Bulb shapes and designations. (*b*) Lamp bases. Designations of International Electrotechnical Commission are shown, where available. (*c*) Filament structures. (*After Illuminating Engineering Society*, *IES Lighting Handbook*, *7th ed*., *1984*)

lamps have a normal life of 750 h, while extended-service and long-life lamps offer 1500–5000 h of life by reducing the efficacy and light output to increase the life (**Table 1**). Users of the longer-life lamps save by buying fewer lamps, but they get much less light and still pay the same amount for electric energy.

The life of lamps is always given on an average basis. Thus, of ten 750-h lamps, two would probably be burned out at 600 h, or 80% life, five after 750 h, or 100% life, and eight after 900 h, or 120% life. The foregoing assumes the lamps operate at rated voltage. The lamp life varies rapidly with voltage. Because utilities cannot supply a constant voltage to all parts of their systems and because there is a variance in wiring from one house to the next, the average life of lamps in homes can vary from the rated value.

Vacuum and gas-filled lamps. Two major classes of lamps are vacuum and gas-filled. Vacuum lamps have had most of the air pumped out of the bulb, and the bulb is sealed. A "getter" chemical in the bulb is then heated inductively, causing it to combine with most of any gas remaining, producing a very high vacuum. Most of the lamps rated 10 W or less are vacuum lamps. Vacuum lamps operate at a lower filament temperature and have a lower efficacy than gas-filled lamps.

Gas-filled lamps have most of the air removed, and the bulb is then filled with an inert gas that acts to slow down the rate at which tungsten molecules evaporate from the filament, enabling the filament temperature to be increased without decreasing the life. The efficacy of the lamp is increased, and it is further increased by coiling the filament in a helix. The effect is enhanced by using heavier gases. The fill gas in standard gas-filled lamps is a mixture of 50–99+% argon (atomic weight 39.91) and nitrogen (at. wt 14.008). The use of the gas krypton (at. wt 82.9) as a fill gas produces an improvement of 7–20% in light output. Krypton lamps are sold at a premium price as energy-saving lamps. Lamps above 75 W are normally gas-filled, while 10–75-W lamps are either vacuum or gas-filled, depending on their design and intended usage.

Color and glare. Vacuum lamps have a higher red radiation and a lower blue radiation than gas-filled lamps (**Fig. 3**). This means that the vacuum lamps appear more red and enhance red objects and darken blue objects more than the same objects when lighted by gas-filled lamps. The daylight lamp has a blue glass bulb which yields light of a color more nearly approaching pure white theoretical emission.

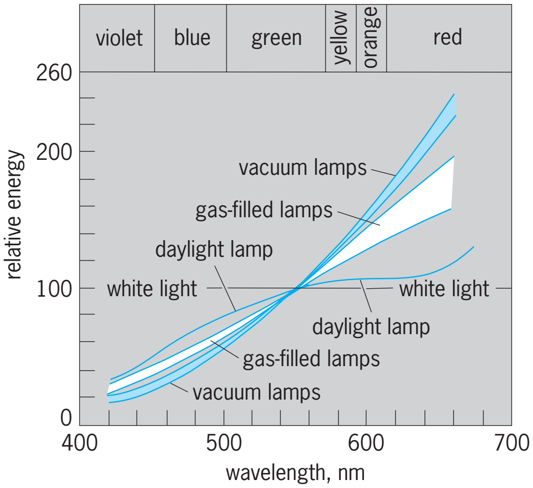


Fig. 3. Spectral energy distribution for important types of incandescent lamps.

TABLE 1. Comparison of life versus light output of 100-W lamps[*]

Lamp type	Life	Initial lumens	% light output
General-service lamp, inside frost	750 h	1740	100
Extended-service lamp, white	2500 h	1460	84
Long-life lamp, inside frost	5000 h	1050	60

[*] Energy consumption of these three lamps is identical.

The color of the light emitted and the glare produced by the bright filament can be controlled by adding a surface treatment or coat to the clear glass bulb. Lamps where a normal filament color is desired, but where glare reduction is needed, are etched or frosted on the inside surface of the glass. The lumen loss is about 0.5% maximum compared to a clear-glass lamp. Where still greater filament hiding is desired, the inside of the glass is coated with a white ceramic powder with a resultant light loss of about 2%. The improved seeing due to a glare reduction far outweighs the light loss, and these are the recommended lamps to use if the lamp is within the field of view.

When light of a particular color such as red, blue, green, or pink is required, the bulbs may be made of colored glass or coated with plastic or ceramic. Some lamps with reflectors built in, such as the PAR (parabolic-aluminized-reflector) lamp (Fig. 2*a*), have dichroic filters for the reflector, so that most of the light passes out the front side of the lamp and the heat, or infrared radiation, passes out the back of the lamp, thus helping to keep the illuminated objects cool. These lamps are often used to illuminate meat cases in stores.

Lumen depreciation. The lumen output of incandescent lamps decreases with age, a characteristic referred to as lumen depreciation. Lumen values stated for lamps are always initial lumens. The rate of depreciation of a lamp is determined by its design and the manner in which it is operated (that is, base-up or base-down, and normal or abnormal voltage). At rated life, most general service lamps emit 75–85% of their initial lumen output.

Effects of voltage variation. Lamps operated at other than rated voltage have their characteristics altered (**Table 2** and **Fig. 4**). The lamp life varies rapidly with voltage, as mentioned above (Fig. 4).

Applications and special types. Incandescent lamps have been developed for many services. Most common are those used in general service and the miniature lamp. Special types have been developed for rough service applications, bake-oven use, severe vibration applications, showcase lamps, multiple lights (three-way lamp), sign lamps, spotlights, floodlights, and insect-control lamps.

Reflector-type lamps are designed with built-in light control, that is, silvering placed on a portion of the lamp's outside surface. The reflecting surface is protected and is effective for the life of the lamp.

Projector-type lamps have molded-glass reflectors, silvered inside the lamp cavity, with either a clear glass cover or a molded control-lens cover. The reflector and the cover are sealed together, forming a lamp with an internal reflector. The parts are of hard glass, and the lamps may be used for outdoor service in floodlight and spotlight installations. This is also the type of lamp used in the sealed-beam headlight for automobiles, locomotives, and airplanes. The contours of the reflector are molded for accurate beam control. With this sturdy structure the filament can be positioned for the best use of the lens, and the lamp has little depreciation during its life. Being constructed of hard glass, it lends itself to high-wattage use.

Miniature incandescent lamps have diverse applications, from the ordinary flashlight to the "grain of wheat" lamps used in surgical and dental instruments. These lamps are designed to give the highest efficiency consistent with the nature of the power source employed.

Special picture-projection lamps are designed for accurate filament location in the focal plane of the optical system, with the filament concentrated as much as possible in a single plane and in a small area. These precision lamps use a prefocus base for accurate positioning of the filament with respect to the base. Projector lamps run at high temperatures, and forced ventilation is frequently required. *See* OPTICAL PROJECTION SYSTEMS.

A special class of lamp is designed for the photographic field, where the chief requirement is actinic quality. Frequently the most important rating is the color temperature, with little regard for economic efficency or life. Photoflood lamps give high illumination for a short life, obtaining twice the lumens of general-service lamps from high filament temperature, and high color temperature with three times the photographic effectiveness. The "daylight blue" photoflood lamp gives a very white light at 4800 K color temperature and 35.8 lm/W.

TABLE 2. Typical performance of large incandescent lamps burned below rated voltage

	% normal operation		
Voltage	Light output	Watts	% initial efficacy
100.0	100.0	100.0	100.0
99.2	97.3	98.8	98.5
98.3	94.4	97.4	96.9
97.5	91.8	96.1	95.5
96.7	89.2	95.0	93.9
95.8	86.4	93.6	92.3
95.0	84.0	92.4	90.9
94.2	81.5	91.2	89.4
93.3	78.0	89.8	86.8
92.5	76.6	88.7	86.3
91.7	74.1	87.5	84.7
90.0	69.5	85.0	81.8
88.3	65.0	82.5	78.8
86.7	60.8	80.3	75.7
85.0	56.6	77.9	72.7
83.3	52.0	75.5	68.8

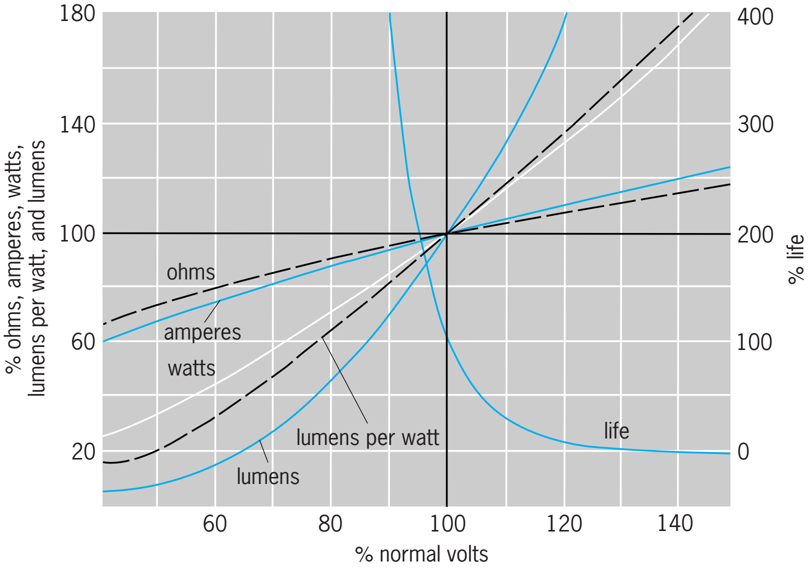


Fig. 4. Characteristic curves of general-service lamps. (*After Illuminating Engineering Society*, *IES Lighting Handbook*, *8th ed.*, *1993*)

Tungsten-halogen lamps. These lamps are made with a fill gas that includes a small amount of one of the halogen elements such as iodine, bromine, or chlorine. The special changes that result from the halogen addition are: (1) the filament temperature can be increased, giving a whiter light output; (2) the depreciation in light output with time is greatly decreased; and (3) the lumen output and the life are increased.

Operation. The filament is enclosed in a small-diameter tubing made of fused quartz instead of glass to withstand the 500°F or 260°C bulb wall temperature required for proper functioning of the halogen gas fill. In operation, as the tungsten evaporates from the filament, it combines with the halogen forming a tungsten-halogen gas. This gas circulates within the lamp, but instead of blackening the bulb wall with tungsten deposits, as happens in regular lamps, the tungsten remains as a gas, until coming in contact with the very high-temperature filament it separates into halogen and tungsten with the tungsten being deposited back on the filament. Dimming the lamp impedes this process and if continued for a long period will result in the blackening of the bulb wall. Occasional operation of the lamp at rated voltage and current will alleviate this problem. This lamp develops a larger amount of ultraviolet radiation than general-service lamps. This may cause problems when lighting objects sensitive to ultraviolet, so precautions may be required.

Configurations. Tungsten-halogen lamps are available in four basic configurations, as replacement lamps, mirrored-reflector (MR) lamps, encapsulated lamps, and tubular lamps.

Replacement lamps are available in sizes of 50 W and less and are used where reflectors are permanently installed. Mirrored-reflector lamps are available in several sizes, with the 2-in.-diameter (5-cm) lamp (MR-16) being the most widely used. They are available in a few different wattages, usually at 12 V, and with a variety of distribution characteristics. They are an excellent choice for illuminating small targets, such as jewelry, to a high illuminance.

Encapsulated lamps, may be used as a substitute for conventional incandescent lamps of the same shape. The tungsten-halogen source is installed (**Fig. 5**) in a conventional lamp shape such as PAR-38, a pressed reflector lamp, 4.75 (38/8) in. (12 cm) in diameter. When replacing a conventional incandescent, a smaller-wattage tungsten-halogen will usually provide equal illuminance.

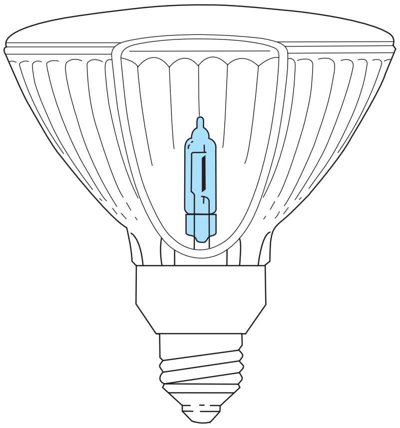


Fig. 5. Encapsulated tungsten-halogen lamp with a section cut away, showing installation of the tungsten-halogen source.

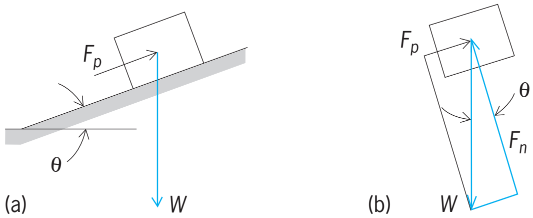
Tubular lamps are available in a double-ended configuration or with a candelabra base. These lamps are available in sizes from 50 to over 500 W. They are commonly used in indirect floor-standing luminaires, wall sconces, and other applications where a compact source size is desirable. When replacing the lamps, touching the tube with the fingers should be avoided, as this may truncate the lamp life.

For other types of incandescent lamps *see* ARC LAMP; INFRARED LAMP. G. R. Peirce; Robert Leroy Smith

Bibliography. Illuminating Engineering Society, *IES Lighting Handbook*, 8th ed., 1993.

Inclined plane

A plane surface inclined at an angle with the line of action of the force that is to be exerted or overcome. The usual application is diagrammed in the **illustration**. A relatively small force acting parallel to the surface holds an object in place, or moves it at a constant speed.



Weight resting on an inclined plane (*a*) with principal forces applied, and (*b*) their resolution into normal force.

In the free-body diagram shown here, three forces act on the object when no friction is present. The forces are its weight W, the force F_p parallel to the surface, and a force F_n normal to the surface. The summation of the forces acting in any direction on a body in static equilibrium equals zero; therefore, the summation of forces parallel to and forces normal to the surface are given by Eqs. (1) and (2). A force

$$F_p - W\sin\theta = 0 \qquad (1)$$

$$F_n - W\cos\theta = 0 \qquad (2)$$

slightly greater than $W\sin\theta$ moves the object up the incline, but the inclined plane supports the greater part of the weight of the object. The principal use of the inclined plane is as ramps for moving goods from one level to another. Wheels may be added to the object to be raised to decrease friction but the principle remains the same. The wedge and screw are closely related to the inclined plane and find wide application. *See* SIMPLE MACHINE; WEDGE. Richard M. Phelan

Incompressible flow

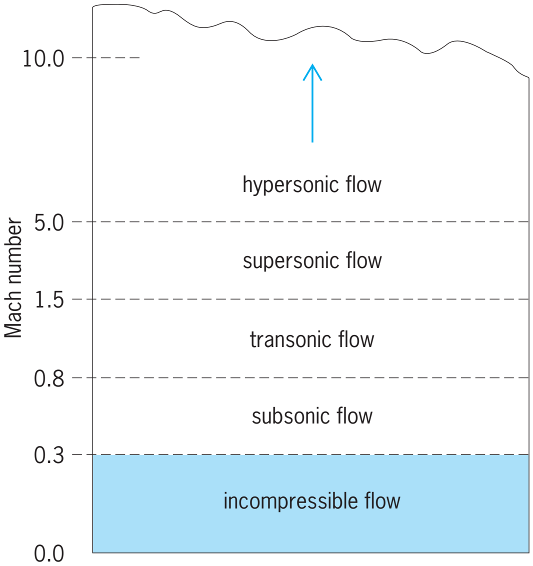
Fluid motion with negligible changes in density. No fluid is truly incompressible, since even liquids can have their density increased through application of sufficient pressure. But density changes in a flow will be negligible if the Mach number, Ma, of the flow is small. This condition for incompressible flow is given by Eq. (1), where V is the fluid velocity and a is the

$$\mathrm{Ma} = \frac{V}{a} < 0.3 \qquad (1)$$

speed of sound of the fluid. It is nearly impossible to attain Ma = 0.3 in liquid flow because of the very high pressures required. Thus liquid flow is incompressible. *See* MACH NUMBER.

Speed ranges in gas dynamics. Gases may easily move at compressible speeds. Doubling the pressure of air—from, say, 1 to 2 atm—may accelerate it to supersonic velocity. In principle, practically any large Mach number may be achieved in gas flow, as shown in the **illustration**. As Mach number increases above 0.3, the four compressible speed ranges occur: subsonic, transonic, supersonic, and hypersonic flow. Each of these has special characteristics and methods of analysis. *See* COMPRESSIBLE FLOW.

Air at 68°F (20°C) has a speed of sound of 760 mi/h (340 m/s). Thus inequality (1) indicates that air flow



Schematic of various speed ranges of fluid motion. Incompressible flow means small but not zero Mach number.

will be incompressible at velocities up to 228 mi/h (102 m/s). This includes a wide variety of practical air flows: ventilation ducts, fans, automobiles, baseball pitches, light aircraft, and wind forces. The result is a wide variety of useful incompressible flow relations applicable to both liquids and gases.

Mass, momentum, and energy relations. The incompressible flow assumption yields very useful simplifications in analysis. For example, a mass balance reduces to a velocity-space relation, independent of fluid density. In steady flow through a variable-area duct, the volume flow Q (m^3/s or ft^3/s) is constant as given in Eq. (2), where V is the average velocity and

$$Q = VA = \text{constant} \qquad (2)$$

A the cross-section area. Constricted areas (nozzles) yield high velocities; for example, in a hypodermic syringe the thumb moves slowly but the jet emerges rapidly from the needle.

The momentum equation for steady, inviscid, incompressible flow reduces to the Bernoulli theorem, given by Eq. (3), where p is pressure, ρ is density

$$\frac{p}{\rho} + \frac{V^2}{2} + gz = \text{constant} \qquad (3)$$

(assumed constant), g is the acceleration of gravity, and z is the elevation. Combining Eqs. (2) and (3) gives, for example, the formula for flow rate of a Bernoulli-type flowmeter. *See* BERNOULLI'S THEOREM; FLOW MEASUREMENT.

Similarly, the general energy equation reduces for incompressible flow to a form of the heat conduction equation, for which many practical solutions are known. *See* CONDUCTION (HEAT); CONVECTION (HEAT).

The incompressible flow assumption also greatly simplifies analysis of viscous flow in boundary layers. Pressure and density are eliminated as variables, and velocity becomes the only unknown, with many accurate solution techniques available. *See* BOUNDARY-LAYER FLOW.

Any disturbance in either a gas or liquid causes the propagation of pressure pulses or “sound waves” in the fluid. This is not a flow but rather a problem in wave propagation and particle oscillation in an otherwise nearly still fluid. *See* SOUND. Frank M. White

Bibliography. I. G. Currie, *Fundamental Mechanics of Fluids*, 2d ed., 1993; V. L. Streeter and E. B. Wylie, *Fluid Mechanics*, 9th ed., 1998; B. Thwaites, *Incompressible Aerodynamics*, 1960, reprint 1987; F. M. White, *Fluid Mechanics*, 4th ed., 1998; C. S. Yih, *Fluid Mechanics*, 1969, reprint 1988.

Index fossil

The skeletal remains (body fossils) or, rarely, the traces of activity (trace fossils) of animals, plants, protistans, and bacteria that can be used to uniquely identify the ages of sedimentary rocks. The term “index fossil” is nearly synonymous with guide fossil or zone fossil. *See* FOSSIL; TRACE FOSSILS.

With the rare exception of fossils that are reworked, that is, eroded from older beds and incorporated into younger sediments, the simple occurrence of particular fossils gives an unambiguous age bracket for the containing strata. Biostratigraphy, the study of the stratigraphic distribution of index or zone fossils, is the primary means by which most sedimentary rocks are assigned to a particular age and then regionally correlated. Index fossils are an important tool in the search for petroleum, coal, and other key geologic resources. *See* GEOLOGIC TIME SCALE; STRATIGRAPHIC NOMENCLATURE; STRATIGRAPHY.

Basis of biostratigraphy. The idea of using index fossils to identify and correlate—or match—strata of a particular geologic age dates to the early 1800s, when the British geologist William Smith established the “law of faunal and floral succession” and used this principle to help trace and map the strata of England, producing the first major geological map. This guiding principle notes that fossil species, genera, and larger groups appear and disappear in the geological record in a definite order or succession (**Fig. 1**); implicit in this notion is the assumption that the first and last appearances of index fossil species are approximately synchronous across large regions.

The “law” of faunal succession makes the tacit assumption that once a species or larger group becomes extinct it can never reappear. The true last appearance of a species, its extinction, is final and global. Thus, species should not be of markedly different ages in different places: finding the species in a particular stratum or horizon at any locality places that stratum into a narrow time slot and, within limits, finding that species in any other location confines that stratum within the same bracket. This concept was developed empirically by Smith and other stratigraphers well before the basic processes of evolution were understood. Darwin's recognition of evolution

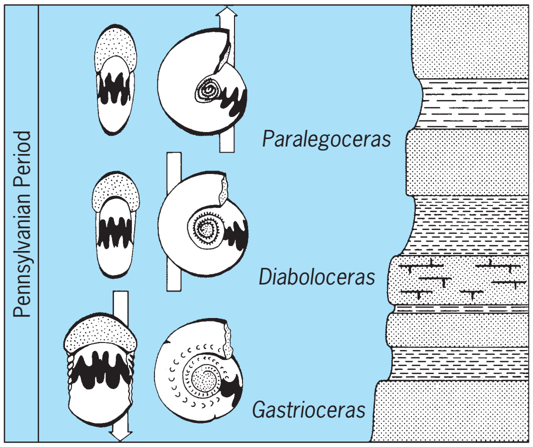


Fig. 1. Three cephalopod genera forming an evolutional sequence. Their short ranges (indicated by bars) and distinctive suture lines (the area between two sutures is shown in black) make them ideal index fossils for the early Pennsylvanian Period. (*After A. K. Miller and W. M. Furnish, Middle Pennsylvanian Schistoceratidae (Ammonoidea), J. Paleontol., 32(2):254, 1958*)

by natural selection provided a theoretical basis for biostratigraphy: evolution is an irreversible process, species arise uniquely from other species at particular times, and once species become extinct they do not reappear later in the geological record; evolution provides an “arrow” of time. Gradually evolving lineages might potentially pose a problem for recognition of index species, raising the question of how to define the time limits of a species that is gradually and continuously evolving. However, the fact is that index fossils are typically recognizable with little or no change for prolonged periods. This empirical observation fostered the concept of punctuated equilibrium. As recognized in the early 1970s by Niles Eldredge and Stephen Jay Gould, evolution commonly does not occur gradually and continuously. Rather, in the model of punctuated equilibrium, most morphological transformation occurs geologically rapidly at the time of speciation—the splitting of new species. Following this burst of change, species lineages may change very little (stasis) for millions of years. Therefore, punctuated equilibrium provides a theoretical basis for the use of index fossils to define biozones. *See* EXTINCTION (BIOLOGY); ORGANIC EVOLUTION; SPECIATION.

The concept of index fossils was originally developed to identify particular stratigraphic units, especially formations, which are basic mappable units. Commonly the formation name was incorporated into the index fossil name: for example, the brachiopod *Gypidula coeymanensis* was identified as unique to the Lower Devonian Coeymans Formation. Paleontologists have compiled lists and catalogs of index fossils and associated formations. A classic example is H. W. Shimer and R. R. Shrock’s *Index Fossils of North America*, which illustrates and lists the ages and formations of occurrence of several thousand species of invertebrate fossils. However, an overreliance on single index fossils led to some errors. There was an implicit assumption that the fossils and the formation that contained them represented the same precise time intervals wherever found. In some instances, the formations represented facies (records of particular environments) that were diachronous, that is, of different ages in different places. If the characteristic index species migrated across environments through time, their appearance in local sections would also be diachronous. In some cases, biostratigraphers tended to split species on very minor taxonomic features that represented slight ecological or local differences, not characteristic of particular ages. Species were assigned ranges coincident with the thicknesses of formations, and too often their actual ranges were not critically evaluated. Also, the application of index fossils in correlation was typically limited by the restriction of those fossils to local areas. *See* FACIES (GEOLOGY).

From the original conception of index fossils evolved the concept of zone fossils. Zone fossils are index fossils that characterize a particular biozone, the fundamental intervals of biostratigraphy; biozones are based on the geological age ranges of particular index fossils (**Fig. 2**). Such zonal fossils ideally should be widespread, abundant, and short-lived (see below). The true range represents the total span of a species from its true first appearance (also called first appearance datum or FAD) to its true last appearance (last appearance datum or LAD). In practice, the local range of a given index fossil is almost never the full range of that species (**Fig. 3**). The local FAD may represent the immigration of the species into the local area in response to favorable environmental conditions; the local LAD (extermination) of the species may record the onset of unfavorable conditions and/or emigration of the species out of the local area. Nonetheless, when gauged against independent indicators of age such as volcanic ash

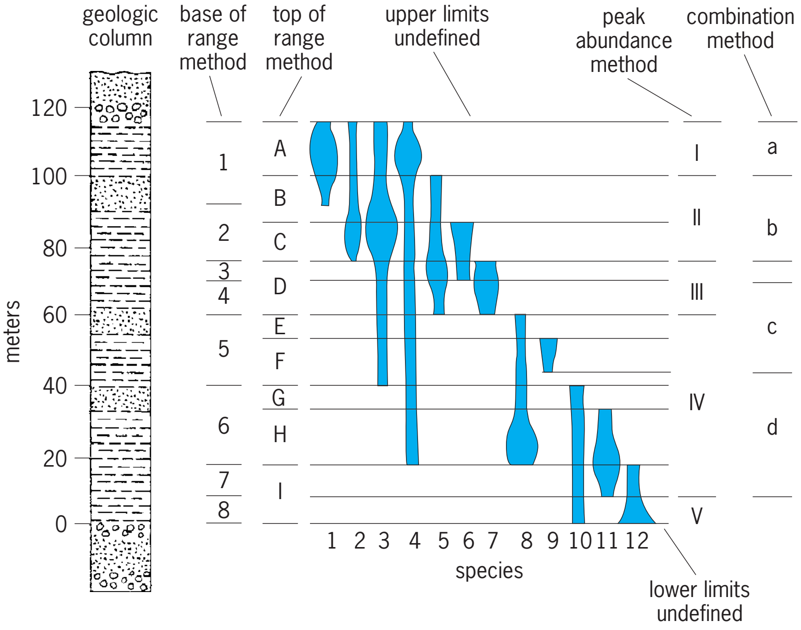


Fig. 2. Different concepts of range zones based on various combinations of overlap of two or more index species. (*From D. R. Prothero and F. Schwab, Sedimentary Geology: An introduction to Sedimentary Rocks, Freeman, 1996*)

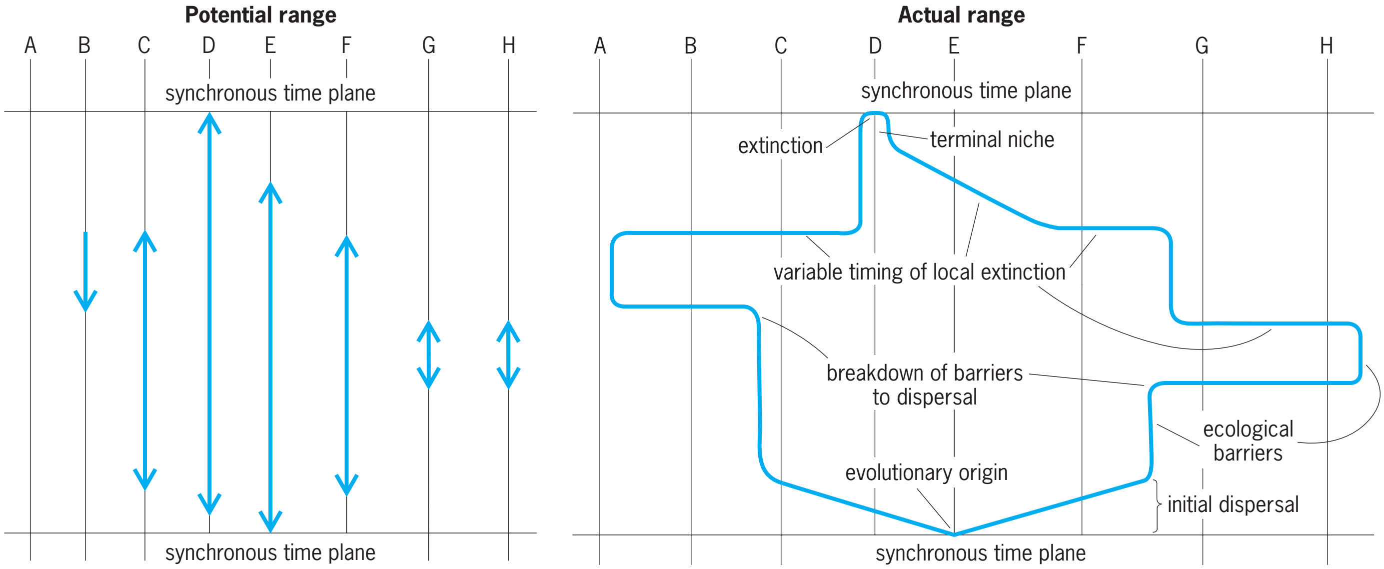


Fig. 3. Potential range versus actual range of an index fossil species. Note that the actual time of occurrence at a given locality is generally some fraction of the actual range of the species because of the limitations of environment, immigration, and emigration. (*From M. Bassett, in D. E. G. Briggs and P. R. Crowther, Palaeobiology: A Synthesis–Part 1, 1991*)

layers or magnetic reversal horizons, many well-chosen index species are demonstrated to have ranges that are roughly synchronous over considerable areas. The biozone is technically defined as a body of strata representing the range of a given species. However, biozones are also a reflection of the time slice representing the duration of the species; this somewhat more abstract concept is referred to as the chronozone.

Biozones are named for their characteristic index fossils, although the names may be abbreviated by just the species name. Separate zonal systems have been developed for different taxonomic groups. For example, the *kockelianus* zone of the Middle Devonian is named for the conodont subspecies *Tortodus kockelianus kockelianus*; the *Agoniatites vanuxemi* Zone is based on a Middle Devonian ammonoid and overlaps in time with the *kockelianus* zone. This system may seem redundant, but it is necessary because various groups evolve at different times and at different rates, and some rocks may carry one fossil and not another. Moreover, the overlap of different index fossils may more precise subdivisions of the rocks. However, calibration among zones is necessary, and generally one zonation scheme, for example, that based on conodonts for Devonian time, is favored for purposes of establishing international stratigraphic boundaries.

In the 1960s, the International Union of Geological Sciences (IUGS) established an International Commission on Stratigraphy and many subcommissions, charged with establishing international boundaries of geological periods and their subdivisions on the basis of biostratigraphy. This commission established the practice of defining particular boundaries, at specific reference sections called global stratotypes, at well-studied, readily accessible localities. At these localities, the boundaries of specific zonal or index species can be precisely located in continuous sections without major gaps. For example, the Silurian-Devonian system boundary was defined precisely at the first appearance (FAD) of the fossil graptolite species *Monograptus uniformis* at the stratotype, a section of alternating shale and limestone in a quarry at Klonk, Czech Republic. Not only was a monument erected at this section in 1972, but a golden spike was driven into the outcrop at the precise point where the lowest specimen of *M. uniformis* was collected, symbolically establishing a standard for global comparison. This notion of the "golden spike" has been figuratively applied in numerous other cases for which global standards were identified based on globally recognizable index fossils at boundary stratotype sections. This provides a needed objectivity to biostratigraphic zonation.

Biostratigraphers long recognized that time increments could be defined more precisely as the overlap of two or more index species. This concept of overlapping of first and/or last appearances is sometimes referred to as Oppel zones in honor of the nineteenth-century paleontologist Carl Albert Oppel, who advocated using overlapping ranges to correlate strata. A more quantitative approach to using index fossils in correlating strata was developed in the 1960s by the American stratigrapher Alan Shaw. Shaw suggested that the ranges of many index fossils in stratigraphic sections to be matched could be plotted against one another as the X/Y axes of a graph. The FADs and LADs of various species, measured in meters (or feet) from the base of each section in each of the two sections, could then be plotted as coordinates of points on the graph (**Fig. 4**). By plotting a series of points, a line of correlation could be determined. Most importantly, the process could be used iteratively to refine the first and last appearances of species on an idealized standard reference section, derived from the most complete and typically thickest stratigraphic section. Such a method is quantitative and makes use of entire assemblages of fossils to define time slices.

Criteria. The identification of fossil species as good index fossils is based on several criteria. In general, the species should represent a group that evolved rapidly. Obviously, the precision with which geologic time can be subdivided is based upon the ranges of the index species: The faster the rate of evolution, the shorter the period of time represented by a given index or zonal fossil species and the narrower the limits of relative age assigned to the rocks containing the species.

An index fossil also must be present in the rocks in sufficient numbers to be found with reasonable

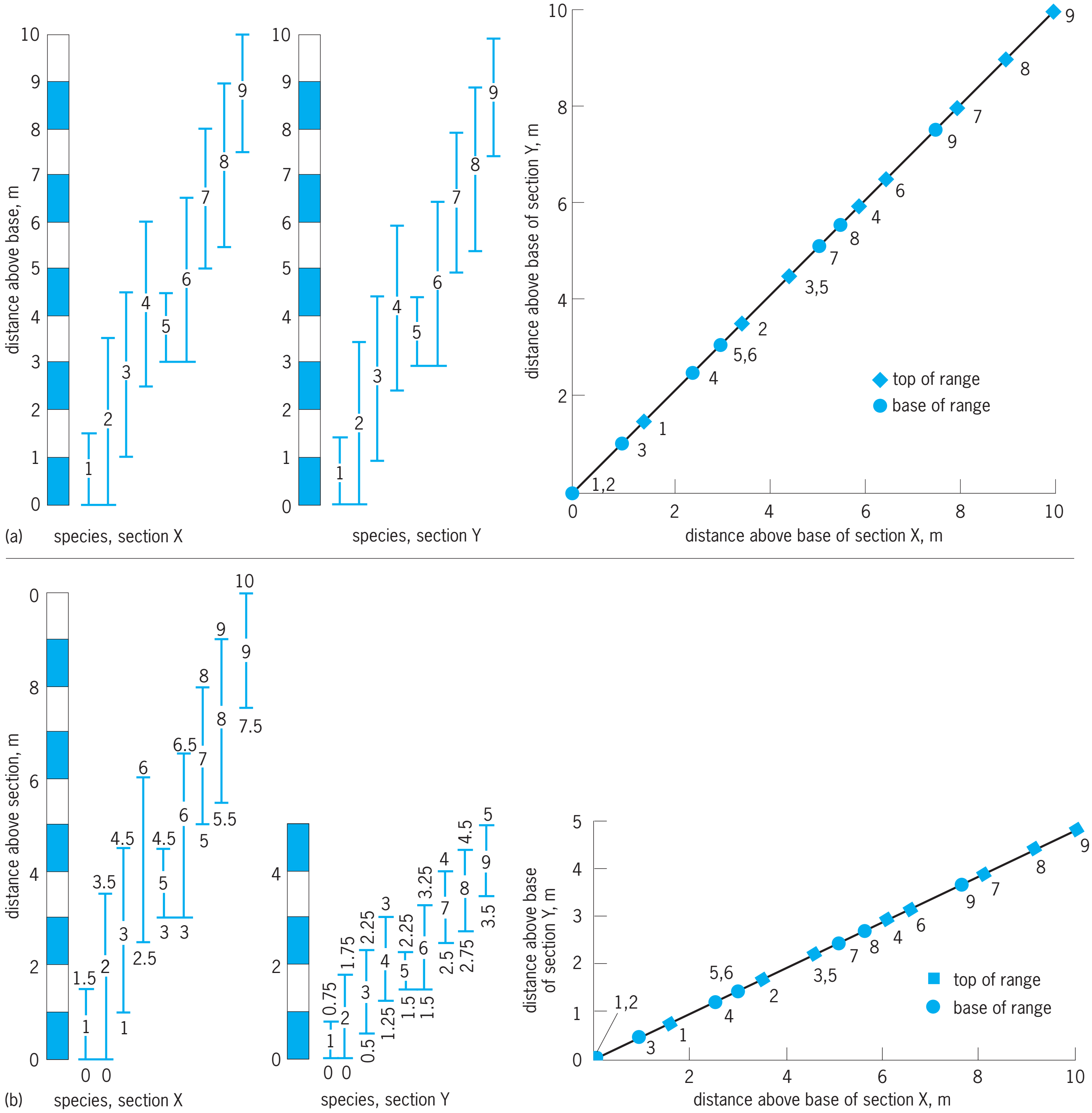


Fig. 4. **Graphic correlation method of Alan Shaw. Note that the *X* and *Y* axes of the graphs are scaled diagrams of measured sections of the two different outcrops to be correlated. The points plotted on the graphs are the corresponding heights of the first or last appearances of various index fossils in each of the stratigraphic sections. The line of correlation (LOC) is a best-fit line drawn through these points. (*a*) Two outcrops with identical stratigraphic sections; note that the heights or first or last appearances of a given species are identical and the angle of the LOC is 45°. (*b*) Correlation plot of two outcrops in which one had a higher rate of sedimentation than the other; note the more tightly spaced FAD and LAD points for the section with lower sedimentation rate. The slope of the LOC is less than 45°. (*From D. R. Prothero and F. Schwab, Sedimentary Geology: An Introduction to Sedimentary Rocks, Freeman, 1996*)**

effort, must be relatively easy to collect and identify, and must be geographically extensive so that the zone it defines is widely recognizable. Ideally, the fossil should be resistant to dissolution or surface weathering. Features that permit the fossil to be identified under conditions of incomplete preservation or weathering are advantageous. Moreover, the usefulness of the fossil is increased if it can be extracted readily from the rock matrix. For example, the minute "teeth" or elements of primitive vertebrates, called conodonts, are composed of apatite (calcium phosphate), which is relatively insoluble in acetic or muriatic acid; these fossils can thus be extracted by dissolving limestones in these acids. Organic-walled fossils, such as pollen, spores, and acritarchs (algal resting cysts), can be released from shales by dissolving the shales in hydrofluoric acid. The study of palynology deals with such organic-walled microfossils (palynomorphs), which are extractable using hydrofluoric acid. Small size in such cases is an advantage as many fossils must be identified from small samples of drill core (material removed from the earth by drilling) or rock chips of drilled well cuttings. Such microfossils, including foraminiferans, radiolarians, conodonts, and spores and pollen, must be studied and identified using microscopy. Micropaleontology is widely used in the petroleum industry. *See* CONODONT; FORAMINIFERA; MICROPALEONTOLOGY; PALYNOLOGY.

The best index fossils would thus be remains of organisms that either live over a wide array of environments or may be distributed widely on water currents or wind. This is most likely to be true for fossils of planktonic (free-floating) organisms, such as many foraminiferans and radiolarians, or nektonic (free-swimming) forms, such as ammonites.

The presence or absence of particular index fossils should not reflect local environments, as the best index fossils are those completely independent of environmental factors. Ideally, index and especially zone fossils should thus occur independent of facies, the records of different environments.

This is not to say that benthic (sea-bottom-dwelling) organisms may not be useful as biostratigraphic markers. Many benthic forms do evolve rapidly and can provide time markers, at least within a particular range of environments. Moreover, many invertebrate animals, even those which are attached in one place (sessile) throughout their adult lives, as is the case for many corals, brachiopods, or clams, have a planktonic or nektonic larval stage in their life cycles. This stage, especially if it involves larvae that feed in the plankton (planktotrophic larvae), may last for many days, and the larvae may thus be dispersed widely on oceanic currents, thereby aiding in the distribution of species over large areas.

Other bottom-dwelling animals are much more tied to specific environmental conditions. Such organisms will typically migrate over time in response to the shifting of their favored environmental conditions, and some may persist for long periods of time with little change. For example, the brachiopod *Lingula* is well adapted to tidal flats and has survived in these environments with little change for more than 500 million years. As such, these fossils may be useful for reconstructing ancient environments and tracing their movements through time; thus they are sometimes termed facies fossils. However, they are not as useful as index fossils, as their occurrence is governed more by environment than by time.

A particularly difficult problem is the correlation of terrestrial and marine environments because few, if any, organisms live in both types of settings. Fortunately, there are some fossils that may enable the connection between these disparate realms. For example, many land plants produce spores and pollen that are adapted for wide dispersal because of their compact, lightweight nature; thus, pollen and spores may be borne on air currents over considerable distances. These palynomorphs may then be transported into and buried within floodplain muds, lake sediments, or even offshore marine sediments. This dispersal potential makes palynomorphs attractive for linking ancient land and sea environments.

Typical index fossils through time. Few types of organisms meet all criteria for an ideal index fossil, and for good reason. Those organisms with widespread or even cosmopolitan (worldwide) distributions and those which are tolerant of a wide variety of environments are also buffered against evolution. A wide distribution and broad environmental tolerance favor survival of species even during times of crisis and major extinction. Thus, widespread species tend to evolve slowly, and vice versa. That said, there are many fossil organisms that may be useful as biostratigraphic markers, at least over particular regions or within particular types of environments. Several groups have proved widely usable as index fossils and have been widely accepted as a basis for global biochronology. Foremost among these forms are acritarchs, spores and pollen, coccoliths, foraminiferans, radiolarians, graptolites, ammonoid cephalopods, and conodonts.

Older Precambrian [prior to 600 million years before present (BP)] sedimentary rocks typically contain few age-diagnostic fossils. During the Archean to middle Proterozoic eons, from about 3.5 to 1.4 billion years ago, most life-forms were simple unicellular organisms such as bacteria. These are often difficult to isolate and identify and have not proved highly useful in biostratigraphy, although there have been attempts to use forms of stromatolites, layered mounds built by cyanobacteria (or blue-green "algae"), to identify particular time increments in the Proterozoic. In the later Proterozoic Eon, from about 1.0 to 0.5 billion years ago (10 to 5 × 10^8 years BP), age-diagnostic fossils—especially the reproductive and resting cysts of marine algae termed acritarchs—are increasingly abundant and provide excellent index fossils for relative dating of sedimentary rocks. In the Phanerozoic Eon, about 5.43 × 10^8 years BP to the present time, a wide array of marine and nonmarine invertebrate and vertebrate animals and plants provide excellent material for subdividing geologic time (see below). Single-celled eukaryotic organisms (cells with nuclei and organelles), such as planktonic algae, first appeared in the early to middle Proterozoic eon about 20 to 15 × 10^8 years BP. These organisms are well represented by acritarchs. Recent advances in the study of acritarchs have established the first detailed biostratigraphic zonation of sediments dating back to at least late Proterozoic time. Acritarchs are also useful microfossils for early Paleozoic biostratigraphy.

The common base of the Phanerozoic Eon, the Paleozoic Era, and the Cambrian Period (about 5.4 × 10^8 years BP) is established as the first appearance of a trace fossil of *Trichophycus pedum*, the oldest complex burrow, in a stratotype section in Newfoundland—one of the rare cases of the use of a trace fossil as an index fossil. Trilobites (extinct, segmented arthropods) provide critical index fossils for zonation in the Cambrian and Ordovician periods. Proteinaceous skeletons of planktonic, colonial organisms called graptolites have proved useful in correlating Ordovician (5.0 to 4.4 × 10^8 years BP) and Silurian (4.4 to 4.17 × 10^8 years BP) rocks. Conodonts, recently recognized as parts of the feeding apparatus of small eellike marine vertebrates, are among the most important index fossils for much of the Paleozoic Era (5.4 to 2.5 × 10^8 years BP) and the Triassic Period (2.5 to 2.0 × 10^8 years BP). Global zonation of the Devonian Period, for example, includes more than 40 successive conodont zones. Ammonoids (coiled cephalopods with variably complex suture patterns reflecting the line of junction of crenulated septa and the external walls of the shells) first appeared in the Devonian Period (about 3.9 × 10^8 years BP) and evolved rapidly throughout their history to final extinction at the end of the Cretaceous Period (6.5 × 10^8 years BP). Goniatite ammonoids with relatively simple zigzag sutures are useful in Late Paleozoic biostratigraphy. Ceratitic ammonoids with intermediate sutural complexity and ammonites with highly complex sutures have proved to be very useful index fossils for the Triassic Period and later Mesozoic Era, respectively. Foraminiferans (protozoans with agglutinated or calcareous tests) evolved in the Cambrian and diversified into many forms during the Phanerozoic Eon; various groups of foraminiferans have been very useful index fossils at different times. Large coiled foraminiferans about the size and shape of rice grains, the fusulinids, are critical index fossils during the Pennsylvanian and Permian periods (3.1 to 2.5 × 10^8 years BP); various smaller benthic foraminiferans are important biostratigraphic microfossils in the Mesozoic (2.5 to 0.65 × 10^8 years BP) and Cenozoic (0.65 to 0 × 10^8 years BP) eras, whereas planktonic foraminiferans are exceptionally important microfossils for correlation in rocks of the late Cretaceous onward. In addition, the beautiful silica lattice skeletons of radiolarians, another planktonic protozoan group, have been used extensively in biostratigraphy of deep-sea sediments throughout the Phanerozoic. From the Cretaceous Period (about 1.2 to 0.65 × 10^8 years BP) onward, the skeletons of planktonic algae become important as index fossils in deep-sea sediments. These microfossils include the tiny diatoms, and skeletal plates of calcite called coccoliths; the latter, which form a dominant component of chalks, belong to a category of ultramicroscopic fossils, sometimes termed nannofossils, that require a scanning electron microscope for study. *See* CEPHALOPODA; FUSULINACEA; GRAPTOLITHINA; RADIOLARIA.

A wide variety of shelly invertebrate groups, including brachiopods, mollusks, corals, ostracodes (tiny bivalved crustaceans), and echinoderms are also used as index fossils for marine deposits. From the Devonian onward, spores and pollen of land plants are useful indices for nonmarine to marginal marine zonation. Freshwater mollusks, ostracodes, and even mammal teeth have also proved useful for zonation of nonmarine sediments in the Cenozoic Era. *See* PALEOECOLOGY; PALEONTOLOGY.

Carlton E. Brett

Bibliography. M. Brasier, *Micropaleontology*, 1986; D. E. G. Briggs and P. R. Crowther, *Palaeobiology: A Synthesis–Part 1*, 1991, Part 2, 2001; E. N. K. Clarkson, *Invertebrate Palaeontology and Evolution*, 3d ed., 1992; D. R. Prothero and F. Schwab, *Sedimentary Geology: An Introduction to Sedimentary Rocks*, Freeman, 1996; H. W. Shimer and R. R. Shrock, *Index Fossils of North America*, 9th printing, MIT Press, 1972.

Indian Ocean

The smallest and geologically the most youthful of the three oceans. It differs from the Pacific and Atlantic oceans in two important aspects. First, it is landlocked in the north, does not extend into the cold climatic regions of the Northern Hemisphere, and consequently it is asymmetrical with regard to its circulation. Second, the wind systems over its equatorial and northern portions change twice each year, causing an almost complete reversal of its circulation.

General Features

The eastern and western boundaries of the Indian Ocean are 147 and 20°E, respectively. In the southeastern Asian waters the boundary is usually placed across Torres Strait, and then from New Guinea along the Lesser Sunda Islands, across Sunda Strait and Singapore Strait. The surface area within these boundaries is about 21% of the surface of all oceans. The average depth is about 12,630 ft (3850 m).

The ocean floor is divided into a number of basins by a system of ridges (**Fig. 1**). The largest is the Mid-Oceanic Ridge, the greater part of which has a rather deep rift valley along its center. It lies like an inverted Y in the central portions of the ocean and ends in the Gulf of Aden. Most of the ocean basins separated by the ridges reach depths in excess of 16,000 ft (5000 m). The Sunda Trench, stretching along Java and Sumatra, is the only deep-sea trench in the Indian Ocean. The Andaman Basin is separated from the open ocean by a 14,600-ft (1400-m) sill. The Red Sea has a maximum depth of 9301 ft (2835 m), but its entrance at the Strait of Bab-el-Mandeb is only 410 ft

Fig. 1. Physiographic features of the Indian Ocean.

(125 m) deep. East of the Mid-Oceanic Ridge, deep-sea sediments are chiefly red clay; in the western half of the ocean, globigerina ooze prevails and, near the Antarctic continent, diatom ooze. *See* BASIN; MARINE GEOLOGY; MARINE SEDIMENTS; MID-OCEANIC RIDGE.

Wind systems. Atmospheric circulation over the northern and equatorial Indian Ocean is characterized by the changing monsoons. In the southern Indian Ocean, atmospheric circulation undergoes only a slight meridional shift during the year (**Fig. 2**). During winter in the Northern Hemisphere, from December through February, the equatorial low-pressure trough is situated at about 10°S and continues into a low-pressure system over northern Australia. A strong high-pressure system lies over Asia. This situation causes the Northeast Monsoon to blow everywhere north of the Equator. The winds cross the Equator from north to south and then usually become northwest winds before reaching the Intertropical Convergence. The subtropical high-pressure ridge of the Southern Hemisphere is situated near 35°S. North of it, the Southeast Trades blow. South of 40°S, winds from the west prevail and are associated with cyclones traveling around Antarctica.

During summer in the Northern Hemisphere, from June through August, a low-pressure system is developed over Asia with the center around Iran. The subtropical high-pressure ridge of the Southern Hemisphere has shifted slightly northward and continues into a high-pressure system over Australia. The Southeast Trades are more strongly developed during this season, cross the Equator from south to north, and become the Southwest Monsoon, bringing rainfall and the wet season to India and Burma. Atmospheric circulation during June through August is much stronger and more consistent than during February. In the Southern Hemisphere the West Wind Belt has shifted about 5° to the north, with westerly winds starting from 30°S and storms becoming stronger and more frequent during winter in that hemisphere. *See* MONSOON METEOROLOGY; TROPICAL METEOROLOGY.

Circulation. The surface circulation is caused largely by winds and changes in response to the wind systems (**Fig. 3**). In addition, strong boundary currents are formed, especially along the western

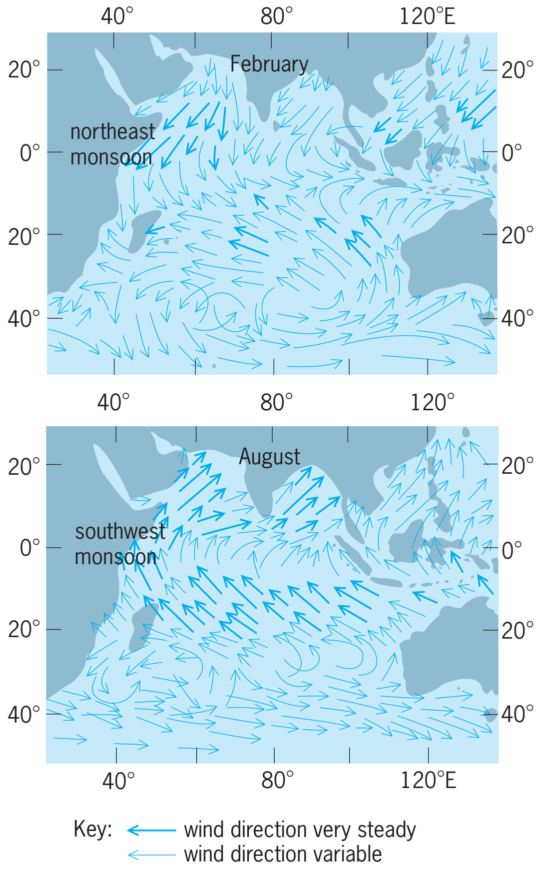


Fig. 2. Winds over the Indian Ocean.

coastline, as an effect of the Earth's rotation and of the boundaries created by the landmasses. In the Southern Hemisphere south of 35°S, a general drift from west to east is found as a result of the prevailing west winds. Near 50°S, where west winds are strongest, the Antarctic Circumpolar Current is embedded in the general West Wind Drift. In the subtropical southern Indian Ocean, circulation is anticyclonic. It consists of the South Equatorial Current flowing west between 10 and 20°S, a flow to the south along the coast of Africa and Madagascar, parts of the West Wind Drift, and flow to the north in the eastern portions of the ocean, especially along the coast of Australia. The flow to the south between Madagascar and Africa is called the Mozambique Current. It continues along the coast of South Africa as the Agulhas Current, most of which turns east into the West Wind Drift. The circulation in the Southern Hemisphere south of 10°S changes only slightly with the seasons.

North of 10°S the changing monsoons cause a complete reversal of surface circulation twice a year. In February, during the Northeast Monsoon, flow north of the Equator is mostly to the west and the North Equatorial Current is well developed. Its water turns south along the coast of Somali and returns to the east as the Equatorial Countercurrent between about 2 and 10°S. In August, during the Southwest Monsoon, the South Equatorial Current extends to the north of 10°S; most of its water turns north along the coast of Somali, forming the strong Somali Current. North of the Equator flow is from west to east and is called the Monsoon Current. Parts of this current turn south along the coast of Sumatra and return to the South Equatorial Current. During the two transition periods between the Northeast and the Southwest monsoons in April-May and in October, a strong jetlike surface current flows along the Equator from west to east in response to the westerly winds during these months.

Although the Antarctic Circumpolar Current reaches to great depths, probably to the bottom, most of the other currents are much shallower. The Agulhas Current and the Somali Current are examples of deep currents, but the remainder of the circulation is limited to the upper surface of the ocean. The deep circulation of the Indian Ocean consists of western boundary currents that carry water of Antarctic origin northward into the various deep-sea basins. This water ascends slowly, is integrated into the surface circulation, and returns south with the Agulhas Current. At the Equator, during the time of the Northeast Monsoon, an Equatorial Undercurrent is found flowing as a subsurface current from west to east. This current is weaker than the corresponding currents in the Pacific and Atlantic oceans. A massive flow of water takes place from the Pacific to the Indian Ocean, called the Indonesian Throughflow. *See* OCEAN CIRCULATION.

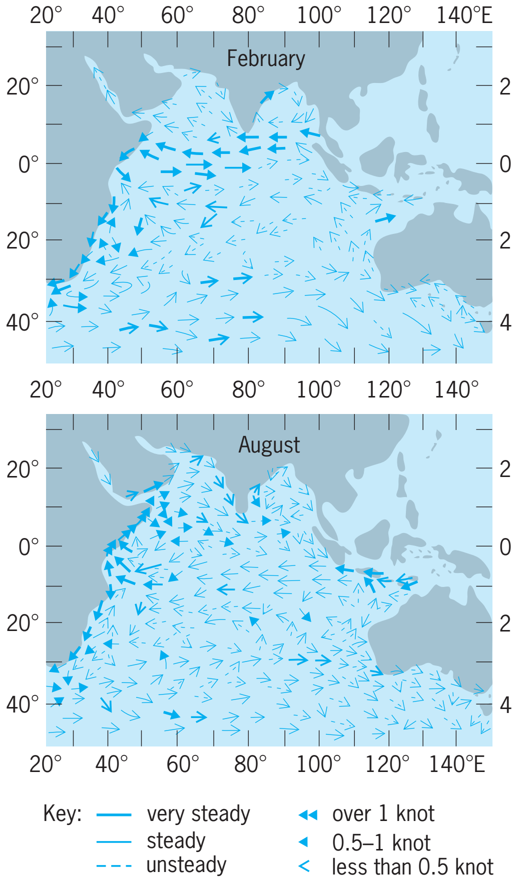


Fig. 3. Surface currents of the Indian Ocean.

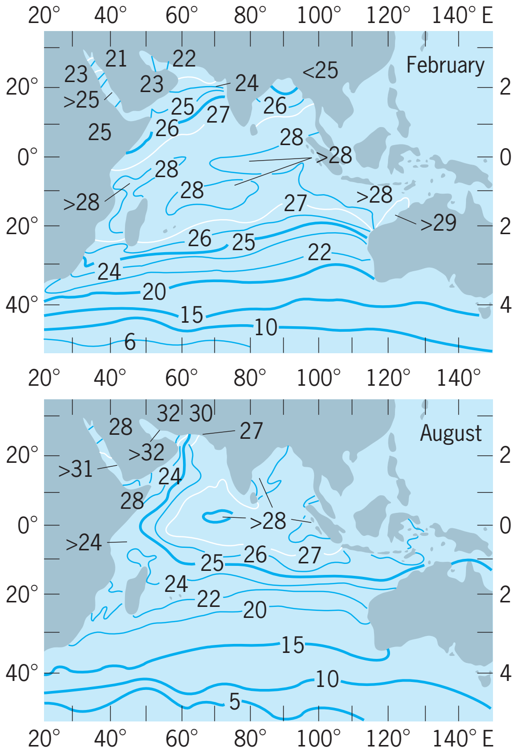


Fig. 4. Sea-surface temperature (°C) of Indian Ocean.

Surface temperature. The pattern of sea-surface temperatures changes considerably with the seasons (**Fig. 4**). During February, when the Intertropical Convergence is near 10°S, the heat equator is also in the Southern Hemisphere and most of the area between the Equator and 20°S has temperatures near 82°F (28°C). The water in the northern parts of the Bay of Bengal and of the Arabian Sea is much cooler, and temperatures below 68°F (20°C) can be found in the northern portions of the Persian Gulf and the Red Sea. In the Southern Hemisphere temperatures decrease gradually from the tropics to the polar regions. Surface circulation affects the distribution of temperature, and warm water spreads south along the coast of Africa and cool water north off the west coast of Australia, causing the isotherms to be inclined from west to east.

During August high temperatures are found in the Northern Hemisphere and in the equatorial region. The Somali Current advects cool water along the coast of Africa to the north. Simultaneously the Southwest Monsoon causes strong upwelling of cooler subsurface water, lowering average water temperature in August to less than 72°F (22°C) along Somalia, and to less than 75°F (24°C) along Arabia. In the Southern Hemisphere isotherms are almost 10° of latitude farther north.

Surface salinity. The distribution of surface salinity is controlled by the difference between evaporation and precipitation and by runoff from the continents (**Fig. 5**).

High surface salinities are found in the subtropical belt of the Southern Hemisphere, where evaporation exceeds rainfall. In contrast, the Antarctic waters are of low salinity because of heavy rainfall and melting ice. Another area of low salinities stretches from the Indonesian waters along 10°S to Madagascar. It is caused by the heavy rainfall in the tropics. The Bay of Bengal has very low salinities, as a result of the runoff from large rivers. In contrast, because of high evaporation the Arabian Sea has extremely high salinities. High salinities are also found in the Persian Gulf and in the Red Sea, representing the arid character of the landmasses surrounding them. The salinity distribution changes relatively little during the year; however, south of India from the Bay of Bengal to the west, a flow of low-salinity water, caused by the North Equatorial Current, can be noticed during February.

Surface water masses. The different climatic conditions over various parts of the Indian Ocean cause the formation of characteristic surface water masses. The Arabian Sea water is of high salinity, has a moderate seasonal temperature variation, and can be classified as subtropical. The water in the Bay of Bengal is of low salinity and always warm, representing tropical surface water. Another type of tropical surface water stretches from the Indonesian waters to the west and is called the Equatorial Surface Water. Subtropical Surface Water has a seasonal temperature variation of between 59 and 77°F (15 and 25°C) and is found in the subtropical regions of the Southern Hemisphere. Its southern boundary is the Subtropical Convergence coinciding with temperatures of about 59°F (15°C). From there, temperature and salinity decrease in the area of the transition water to

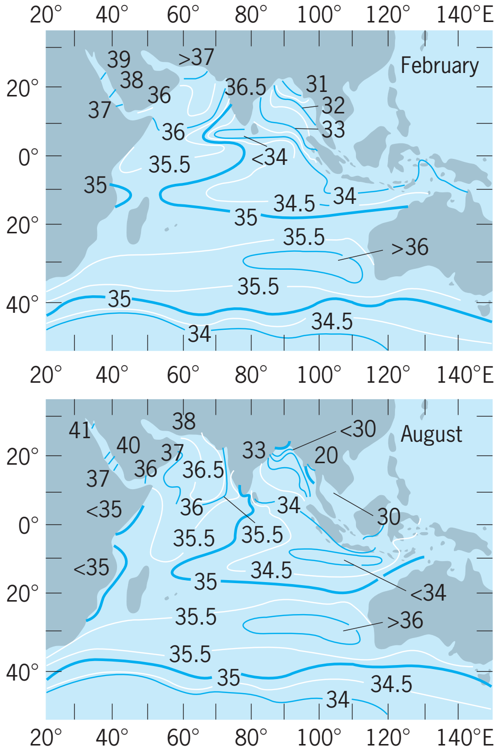


Fig. 5. Sea-surface salinity of the Indian Ocean in parts per thousand by weight (‰).

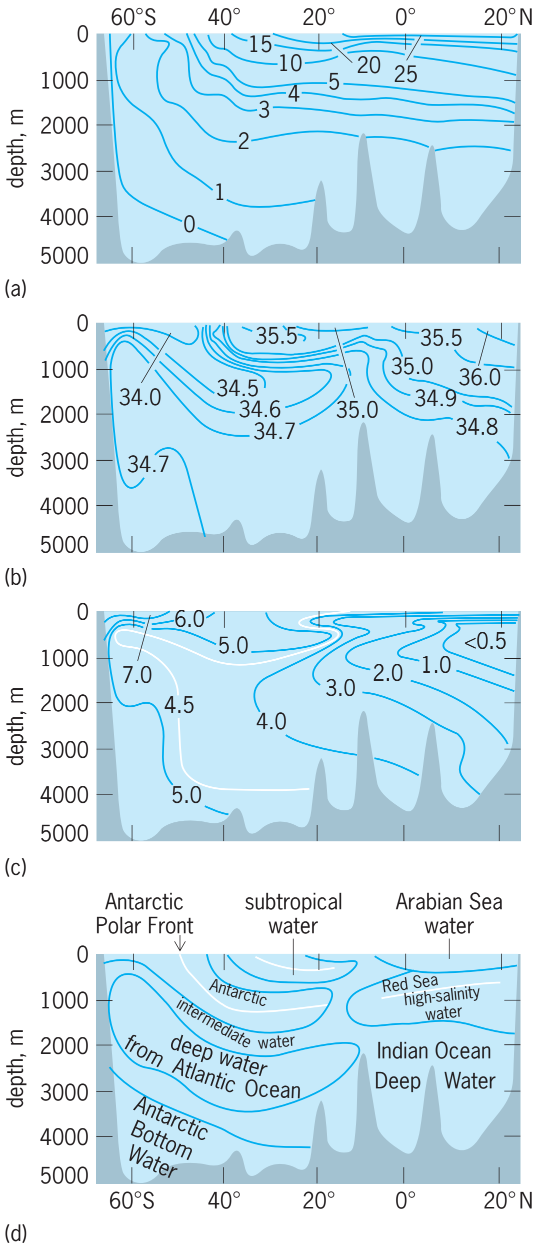


Fig. 6. Indian Ocean from the Arabian Sea to Antarctica showing (*a*) temperature (°C), (*b*) salinity (‰), (*c*) oxygen content (ml/liter), and (*d*) water masses.

about 39°F (4°C) at the Antarctic Polar Front. South of the Antarctic Polar Front, Antarctic Surface Water of low salinity is found; its temperature is near the freezing point in winter and is approximately 36°F (2°C) in summer.

Subsurface water masses. In the Indian Ocean, warm water of more than 59°F (15°C) occupies only a very thin surface layer (**Fig. 6*a***). This layer contains the tropical and subtropical water masses. Subtropical water of the Southern Hemisphere spreads as a subsurface salinity maximum toward the Equator at depths between 330 and 660 ft (100 and 200 m) beneath the lighter tropical surface water.

Water of high salinity formed in the Red Sea and in the Persian Gulf leaves these basins and spreads as a subsurface layer of high salinity throughout the Arabian Sea (Fig. 6*b*). To the south it can be traced as far as Madagascar, and to the east as far as Sumatra. Water of low salinity and a temperature of approximately 39°F (4°C) sinks near the Antarctic Polar Front and spreads north as a salinity minimum. It is called the Antarctic Intermediate Water.

The deep and bottom water of the Indian Ocean is of external origin. South of Africa, water of rather high salinity, originating south of Greenland in the North Atlantic Ocean, enters the Indian Ocean, filling the layers between 4900- and 9800-ft (1500- and 3000-m) depths. The water below the 9800-ft (3000-m) depth is Antarctic Bottom Water originating in the Weddell Sea. Its temperature is lower than 32°F (0°C).

Because of their origin at the sea surface, both the Antarctic Intermediate Water and the Antarctic Bottom Water are of rather high oxygen content (Fig. 6*c*). Since their residence time is long, the water masses of the northern Indian Ocean below the surface layer have a very low oxygen content. An oxygen minimum is associated with the Red Sea water and extends south between the Antarctic Intermediate Water and the North Atlantic Deep Water. *See* ANTARCTIC OCEAN; RED SEA; SEAWATER.

Upwelling. Several areas of upwelling are found along the shores of the Indian Ocean, but all of them are seasonal in character, in contrast to upwelling areas in the other oceans. They are especially developed during the Southwest Monsoon season from May through September, when upwelling takes place in the Banda Sea, south of Java, along the coast of Somali, and off Arabia.

The strongest upwelling is found along the Somaliland coast, where surface temperatures during this season may be as low as 57°F (14°C). Water of high nutrient content ascends from deep levels and is integrated into the surface flow. These areas have a high biological production. During the Northeast Monsoon, some upwelling takes place in the Andaman Sea and off the west coast of India. *See* UPWELLING.

Tides. Both semidiurnal and diurnal tides occur in the Indian Ocean. The semidiurnal tides rotate around three amphidromic points situated in the Arabian Sea, southeast of Madagascar, and west of Perth. The diurnal tide also has three amphidromic points: south of India, in the Mozambique Channel, and between Africa and Antarctica. It has more the character of a standing wave, oscillating between the central portions of the Indian Ocean, the Arabian Sea, and the waters between Australia and Antarctica.

Along the coast of Africa, in the Bay of Bengal, and along the northwest coast of Australia semidiurnal tides prevail, with two high waters each day. Mixed tides are found around the Arabian Sea and along the coasts of Sumatra and Java. Prevailing diurnal tides with only one high water each day occur only along southwest Australia. The ranges of spring tides are high in the Arabian Sea, for example, at Aden and Bombay. The Bay of Bengal has high tidal ranges, chiefly along the coast of Burma, particularly at Rangoon. Moderately high tides are found along

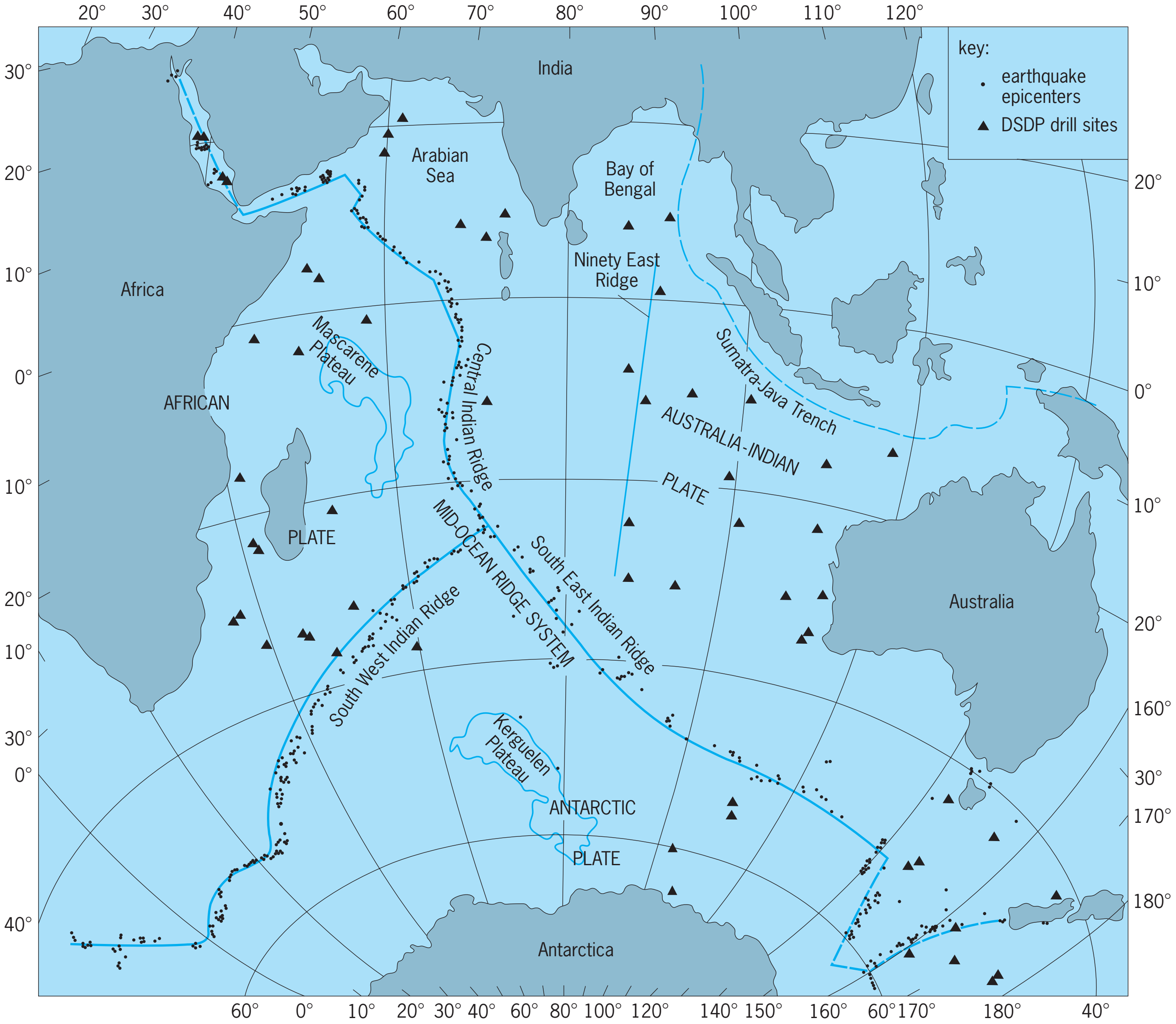


Fig. 7. Mid-ocean ridge system of the Indian Ocean. DSDP = Deep Sea Drilling Project.

the coast of Sumatra and Java, but spring tides along the northwest coast of Australia are much higher. *See* TIDE.

Klaus Wyrtki

Geology and Geophysics

The Indian Ocean, the most youthful of the major ocean basins, was probably formed entirely within the last 1 × 10^8 years. This ocean, in effect, represents that area created during the fragmentation (through processes of sea-floor spreading and plate tectonics) of the ancient supercontinent of Gondwanaland. *See* CONTINENTAL DRIFT; JURASSIC; PLATE TECTONICS.

Major plates. The Indian Ocean is composed of parts of three major plates: on the west, the African plate; on the east and north, the Australian-Indian plate; and on the south, the Antarctic plate. There are many spectacular morphologic features within the Indian Ocean. The major one is the system of mid-ocean ridges that consists of three branches (**Fig. 7**). The mid-ocean ridge system is seismically active and marks the sites of the creation of new oceanic crust. Between the mid-ocean ridges and the surrounding continents lie the major basins of the Indian Ocean. With one exception, the boundaries of the Indian Ocean are characterized by passive continental margins, that is, margins that are not seismically active and that lie well away from the boundaries between major plates. *See* CONTINENTAL MARGIN.

In addition to the mid-ocean ridge systems, ocean basins, and continental margins, the Indian Ocean contains several relatively shallow ridges and plateaus that are free from earthquake activity and whose origins remain enigmatic. One prominent example of these features is the Ninety East Ridge, a 2400-mi-long (4000-km), spectacularly linear feature which may represent a scar left in the ocean floor in the wake of India's northward drift before it collided with the Asian mainland. Another major topographic plateau is the Mascarene Plateau. It lies to the northeast of Madagascar and includes the islands of the Seychelles and Mauritius. Although the origin of the entire plateau remains controversial, there is general agreement that the Seychelles Islands represent a microcontinental fragment that was left behind when the major continents were dispersed. Lying well to the south is the Kerguelen Plateau, whose origin is also unknown. The available data, however, suggest that at least the northern portion of the plateau was formed by submarine volcanism and uplift within a previously existing ocean basin.

The northeastern boundary of the Indian Ocean is clearly defined by an active arc-trench system lying south of Sumatra and Java. This trench system marks the locus of the destruction of Indian Ocean crust that is being transported northward (part of the Australian-Indian plate) as the result of active sea-floor spreading along the Southeast Indian Ocean Ridge.

Distribution of sediments. The distribution of sediments within the Indian Ocean is controlled primarily by three factors: the age of the underlying sea floor, the proximity of the continents, and the barrier effects of the major ridges that are so prevalent throughout the Indian Ocean. Thick sedimentary deposits characterize those parts of the ocean lying near the continental margins. In particular, the major drainage basins of eastern and western India bring immense quantities of sediment from the Himalayas to the northern Indian Ocean floor via the Ganges and Indus rivers. Hence, great thicknesses of sediment are found in the Bay of Bengal and in the Arabian Sea, where accumulations exceed several miles.

Fig. 8. Geotectonic image of the Indian Ocean as reflected in the satellite-derived gravity field of the region. A few specific structural features are identified as follows: Spreading centers: A, Carlsberg Ridge; B, Southwest Indian Ridge; C, Southeast Indian Ridge. Fracture zones: D, Indian-Antarctic; E, Ninety East Ridge (tentatively identified as a fracture zone). Plateaus: F, Kerguelen; G, Chagos-Lacadive; H, Mascarene; J, Agulhas. Zone of intraplate deformation, K. (*W. Haxby*, *Lamont-Doherty Geological Observatory of Columbia University*)

The active mid-ocean ridge systems, because of their relative youth, elevation, and great distance from terrigenous sediment sources, mark the sites of little or no sediment accumulation.

No drill holes and very few near-surface sediment samples exist for the high southern latitudes between about 0° longitude and 100°E longitude. Samples from other parts of the circum-Antarctic Ocean region indicate the dominance of sediments composed of the remains of silicic organisms. Therefore, it is expected that similar sediments characterize the extreme Southwest Indian Ocean. As the Antarctic continental margin is approached, significant quantities of terrigenous materials derived from the continent are also expected.

There has been much seismic research activity focused on the Indian Ocean segment of the Antarctic continental margins. While these studies are designed primarily to obtain more information about the early geologic history of the region, they are also of great interest to various national groups who would like to better assess the hydrocarbon potential of the offshore Antarctic.

Magnetic reversals. The study of magnetic reversals and the lineation stripes formed in association with the spreading mid-ocean ridge system has provided the primary tool for dating the underlying crust in the Indian Ocean. These data are also used to determine the time at which the various continental fragments were split away from Antarctica, and the rates and directions they migrated. The magnetic lineation data are rather convincing, and their interpretations are generally accepted for the southeast and northwest branches of the Indian Ocean mid-ocean system. However, the section of the mid-ocean ridge lying between Africa and Antarctica, the Southwest Indian Ridge, is characterized by numerous closely spaced offsets. This fact, coupled with a paucity of data, leaves much of the Southwest Indian Ocean still poorly understood in terms of its sea-floor spreading history. *See* PALEOMAGNETISM.

Drilling programs. An extensive program of drilling carried out as part of the Deep Sea Drilling Project (DSDP) (Fig. 7) confirmed the predicted ages of the sea floor obtained from the analysis of marine magnetic anomalies. They also contributed to present understanding of the plate tectonic evolution, the paleoenvironment, and the paleocirculation patterns of the Southeast Indian Ocean.

History. Unfortunately, major portions of the high-latitude regions of the Indian Ocean, particularly the region between Africa and Antarctica, remain poorly explored. Although there is general agreement about the reconstruction of major parts of the continental masses surrounding the Indian Ocean to a prerifted Gondwanaland configuration, the history prior to about 1 × 10^8 years remains controversial.

It has been postulated that the fragmentation of the Gondwanaland continent, especially the final departure of Australia from Antarctica in the early Cenozoic, allowed the formation of a major oceanic circulation pattern around the continent of Antarctica. Paleomagnetic evidence clearly indicates that the Antarctic continent resided in a near-polar position long before it became glaciated; hence an additional factor or condition was needed to cause glaciation. The removal of major topographic barriers by sea-floor spreading allowed the circumpolar current pattern to form and led to complex processes of latitudinal heat exchange. This may have provided the triggering mechanism that led to the eventual glaciation of the Antarctic continent about 2.5 × 10^7 years ago. *See* CENOZOIC.

Mapping the tectonic features of the ocean basins is possible by utilizing minute variations in the average height of the sea surface as measured from satellites; thus, a much more detailed map of the ocean gravity field can be deduced. The gravity field in turn reflects the subtle lateral density variations caused by the sea-floor topography and also by tectonic structures that are buried. The satellite observations have thus led to a synthetic representation of sea-floor relief, including fracture zone "traces," which reveal the pathways along which the Gondwanaland continents drifted apart (**Fig. 8**). The potential for reconstructing the details of the formation of the Indian Ocean sea floor has been greatly enhanced.

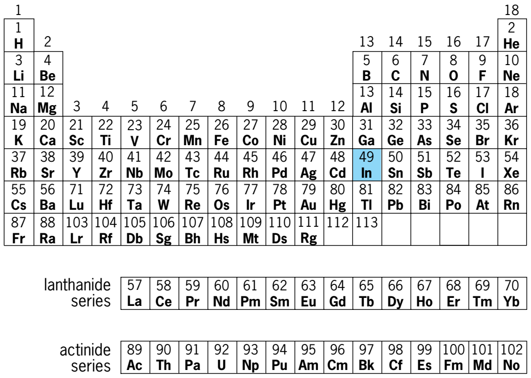
Geotectonic imagery has shown that the trends of the fracture zones in the northwest Indian Ocean are not parallel to the Ninety East Ridge (Fig. 7), thus making it unlikely that this ridge represents the trace of an old fracture zone, a previously popular hypothesis. Also, in the area south of India but north of the mid-ocean ridge system, both the satellite data and other geophysical data collected from ships indicate that the Indian plate is undergoing major compressional deformation within its boundaries. Continued deformation there could eventually lead to the formation of a new collision plate boundary and the partition of the Indian plate into two plates. It is thus clear that a knowledge of the detailed tectonic history of the region can have an important impact on such far-ranging topics as climate evolution, ocean circulation, and mineral resource predictions. *See* PALEOCEANOGRAPHY.

Dennis E. Hayes

Bibliography. S. Chandra, B. Arunachalam, and V. Suryanarayan (eds.), *The Indian Ocean and Its Islands: Strategic, Scientific, and Historical Perspectives*, 1993; D. N. Desai (ed.), *Oceanography of the Indian Ocean*, 1992; C. A. Geller, J. K. Weissel, and R. N. Anderson, Heat transfer and intraplate deformation in the central Indian Ocean, *J. Geophys. Res.*, 88(B2):1018–1032, 1983; S. Hastenrath, P. J. Lamb, and L. L. Greischer, *The Climatic Atlas of the Indian Ocean*, 3 vols., 1989; W. F. Haxby, Geotectonic imagery: Quantum leap in mapping power, *Lamont Newsl.*, Fall 1982; K. Wyrtki, *Oceanographic Atlas of the International Indian Ocean Expedition*, 1989.

Indium

A chemical element, In, atomic number 49, a member of group 13 and the fifth period of the periodic table. Indium has a relative atomic weight of 114.82. Indium occurs in the Earth's crust to the extent of about 0.000001% and is normally found in concentrations of 0.1% or less. It is widely distributed in many ores and minerals but is largely recovered from the flue dusts and residues of zinc-processing operations. *See* PERIODIC TABLE.



1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1 H																	2 He
3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
11 Na	12 Mg											13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
55 Cs	56 Ba	71 Lu	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
87 Fr	88 Ra	103 Lr	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112	113					

lanthanide series	57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb
actinide series	89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No

Indium is used in soldering for electronics and as a component of the intermetallic semiconductor used for germanium transistors. Indium arsenide, antimonide, and phosphide are semiconductors with unique properties. Other uses of indium are sleeve-type bearings to reduce corrosion and wear, glass-sealing alloys, and dental alloys. *See* GERMANIUM.

Edwin M. Larsen

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; A. J. Downs (ed.), *Chemistry of Aluminium, Gallium, Indium and Thallium*, 1993; W. F. Gale and T. C. Totemeier, *Smithells Metals Reference Book*, 8th ed., 2004; D. R. Lide, *CRC Handbook Chemistry and Physics*, 85th ed., CRC Press, 2004; K. T. Tsen (ed.), *Ultrafast Physical Processes in Semiconductors: Semiconductors and Semimetals*, vol. 67, 2000.

Indole

The parent compound of a group of organic heterocyclic compounds containing the indole nucleus, which is a benzene ring fused to a pyrrole ring as in indole itself. Indole can exist in two tautomeric forms, the more stable enamine form and the

3-H-indole or imine form. Unsubstituted 3-H-indoles (sometimes called indolenines) and a structural isomer of indole, isoindole, are not stable, but have been shown to be reaction intermediates. They are isolable when properly substituted. *See* PYRROLE.

The importance of the indole ring lies in its presence in a large number of naturally occurring compounds. Some of these are the plant growth hormone indole-3-acetic acid, the animal tissue constitutent serotonin, the amino acid tryptophan, the pigment indigo, and the antitumor antibiotic mitomycin A. In addition, the indole nucleus is found in the indole alkaloids and in a number of synthetic drugs and pigments.

Physical properties. Indole is a steam-volatile, colorless solid, melting point 52.5°C (126°F), boiling point 253°C (487°F). It is found in small amounts in coal tar, feces, and flower oils. Indole has a characteristic ultraviolet spectrum with absorption maxima at 216, 266, 287, and 276 nanometers. It fluoresces (shines) when irradiated with ultraviolet light. These, and other spectral properties, vary widely for the many derivatives of the enamine form.

Preparations. Indoles are generally synthesized by constructing a five-membered nitrogen ring on a preformed benzene ring. The Fischer synthesis is the most common and versatile of many known methods. In this method, the phenylhydrazone of some ketone is treated with acid to give the substituted indole. The R groups in the ketone may be varied widely, and the phenylhydrazone ring may be substituted. Other syntheses include the Bischler synthesis from anilines and haloketones and the ring closure of *o*-toluides.

Chemical properties. Despite the presence of nitrogen, indole is not basic in the sense that it dissolves in aqueous acid or turns litmus blue. In fact, the hydrogen on the nitrogen is about as acidic as an aliphatic alcohol hydrogen ($pK_a =$ about 17). Indole is an aromatic compound and undergoes electrophilic substitutions much like benzene, although it is much more reactive than benzene. Its reactivity is comparable to that of phenol, and it undergoes a number of reactions similar to those of phenol. Much of the chemistry of indole can be summarized mechanistically by reaction scheme. Essentially, indole is an efficient provider of electrons which can react with an electrophile (E^+) to yield a substituted product, or which can take part in the displacement of a nucleophile (Nu^-) from carbon to yield an alkylated product. When the electrophile is a polarized double bond such as a carbonyl or an imine, addition reactions take place.

Alkylation and acylation are more facile when the nucleophilic character of the indole is enhanced by converting it into a metallic salt, using the acidic proton on nitrogen. Thus, the nitrogen Grignard reagent of indole can be prepared from and any simple Grignard reagent. The metallic salt or complex is then alkylated or acylated by a nucleophilic displacement. Both the N- and the 3-substituted products are formed.

In the third type of reaction, indole reacts with aldehydes or ketones, with imines, and with activated alkenes.

Indole itself reacts slowly with air and rapidly with most oxidizing agents to give intractable polymeric tars. However, substituted indoles undergo a number of smooth oxidations to give isolable products. Air oxidations appear to involve a peroxide intermediate formed from 1,2,3,4-tetrahydrocarbazole. The peroxide can then undergo a number of reactions. Indole can be reduced to give 2,3-dihydroindoles (indolines) or octahydro derivatives.

Indole alkaloids. The indole alkaloids are a large group of substances containing the indole nucleus and can be isolated from plants. They contain a number of physiologically active materials, such as strychnine, reserpine, some forms of curare poison, and the rye-fungus drug, ergot. The indole alkaloids are synthesized in tissue from the amino acid tryptophan by a series of complex enzyme-moderated processes. The processes involve the decarboxylation of to give tryptamine, which reacts with some aldehyde to give a compound containing the B-carboline ring system. When the R on the aldehyde is methyl, the simple B-carboline alkaloids, such as harman, result. When R is a more complex fragment, the more complex indole alkaloids such as strychnine and reserpine result. When tryptamine is combined with a five-carbon unit, a key component of ergot, lysergic acid, arises. When the acid group of lysergic acid is converted to the *N*,*N*-diethylamide, lysergic acid diethylamide (LSD;) is obtained. LSD is a synthetic, and not a naturally occurring, substance. There are many hundreds of other indole alkaloids. *See* ALKALOID; HETEROCYCLIC COMPOUNDS. James M. Bobbitt

Bibliography. Atta-ur-Rahman and A. Basha, *Indole Alkaloids*, 1998; J. A. Joule and K. Mills, *Heterocyclic Chemistry*, 4th ed., 2000; R. J. Sundberg, *Indoles*, 1996; K. P. C. Vollhardt and N. E. Schore, *Organic Chemistry*, 5th ed., 2006.

Inductance

That property of an electric circuit or of two neighboring circuits whereby an electromotive force is induced (by the process of electromagnetic induction) in one of the circuits by a change of current in either of them. The term inductance coil is sometimes used as a synonym for inductor, a device possessing the property of inductance. *See* ELECTROMAGNETIC INDUCTION; ELECTROMOTIVE FORCE (EMF); INDUCTOR.

Self-inductance. For a given coil, the ratio of the electromotive force of induction to the rate of change of current in the coil is called the self-inductance L of the coil, given in Eq. (1), where $\mathcal{E}$ is the elec-

$$L = -\frac{\mathcal{E}}{dI/dt} \qquad (1)$$

tromotive force at any instant and dI/dt is the rate of change of the current at that instant. The negative sign indicates that the induced electromotive force is opposite in direction to the current when the current is increasing (*dI/dt* positive) and in the same direction as the current when the current is decreasing (*dI/dt* negative). The self-inductance is in henrys when the electromotive force is in volts, and the rate of change of current is in amperes per second.

An alternative definition of self-inductance is the number of flux linkages per unit current. Flux linkage is the product of the flux Φ and the number of turns in the coil N. Then Eq. (2) holds. Both sides of

$$L = \frac{N\Phi}{I} \tag{2}$$

Eq. (2) may be multiplied by I to obtain Eq. (3), which may be differentiated with respect to t, as in Eqs. (4).

$$LI = N\Phi \tag{3}$$

$$\begin{aligned} L\frac{dI}{dt} &= N\frac{d\Phi}{dt} = -\mathcal{E} \\ L &= -\frac{\mathcal{E}}{dI/dt} \end{aligned} \tag{4}$$

Hence the second definition is equivalent to the first.

Self-inductance does not affect a circuit in which the current is unchanging; however, it is of great importance when there is a changing current, since there is an induced emf during the time that the change takes place. For example, in an alternating-current circuit, the current is constantly changing and the inductance is an important factor. Also, in transient phenomena at the beginning or end of a steady unidirectional current, the self-inductance plays a part. *See* ELECTRIC TRANSIENT.

Consider a circuit of resistance R and inductance L connected in series to a constant source of potential difference V. The current in the circuit does not reach a final steady value instantly, but rises toward the final value $I = V/R$ in a manner that depends upon R and L. At every instant after the switch is closed the applied potential difference is the sum of the iR drop in potential and the back emf $L\, di/dt$, as in Eq. (5), where i is the instantaneous value of the current. Separating the variables i and t, one obtains Eq. (6). The solution of Eq. (6) is given in Eq. (7).

$$V = iR + L\frac{di}{dt} \tag{5}$$

$$\frac{di}{(V/R) - i} = \frac{R}{L}\,dt \tag{6}$$

$$i = \frac{V}{R}(1 - e^{-(R/T)t}) \tag{7}$$

The current rises exponentially to a final steady value V/R. The rate of growth is rapid at first, then less and less rapid as the current approaches the final value.

The power p supplied to the circuit at every instant during the rise of current is given by Eq. (8).

$$p = iV = i^2R + Li\frac{di}{dt} \tag{8}$$

The first term i^2R is the power that goes into heating the circuit. The second term $Li\, di/dt$ is the power that goes into building up the magnetic field in the inductor. The total energy W used in building up the magnetic field is given by Eq. (9). The energy used in

$$\begin{aligned} W &= \int_0^t p\,dt = \int_0^t Li\,\frac{di}{dt}\,dt \\ &= \int_0^I Li\,di = {}^1\!/_2 LI^2 \end{aligned} \tag{9}$$

building up the magnetic field remains as energy of the magnetic field. When the switch is opened, the magnetic field collapses and the energy of the field is returned to the circuit, resulting in an induced emf. The arc that is often seen when a switch is opened is a result of this emf, and the energy to maintain the arc is supplied by the decreasing magnetic field.

Mutual inductance. The mutual inductance M of two neighboring circuits A and B is defined as the ratio of the emf induced in one circuit $\mathcal{E}$ to the rate of change of current in the other circuit, as in Eq. (10).

$$M = \frac{\mathcal{E}_B}{-(dI/dt)_A} \tag{10}$$

The SI unit of mutual inductance is the henry, the same as the unit of self-inductance. The same value is obtained for a pair of coils, regardless of which coil is taken as the starting point.

The mutual inductance of two circuits may also be expressed as the ratio of the flux linkages produced in circuit B by the current in circuit A to the current in circuit A. If Φ_{BA} is the flux threading B as a result of the current in circuit A, Eqs. (11) hold. Integration leads to Eq. (12).

$$\mathcal{E}_B = -N_B\frac{d\Phi_{BA}}{dt} = -M\frac{dI_A}{dt} \tag{11}$$

or

$$N_B\, d\Phi_{BA} = M\, DI_A$$

$$M = \frac{N_B\Phi_{BA}}{I_A} \tag{12}$$

See INDUCTANCE MEASUREMENT. **Kenneth V. Manning**

Bibliography. B. I. Bleaney and B. Bleaney, *Electricity and Magnetism,* 3d ed., 1976, reprint 1989; E. M. Purcell, *Electricity and Magnetism,* 2d ed., 1985.

Inductance measurement

The measurement of self- or mutual inductance. An electrical reactance such as the angular frequency ($2\pi f$, where f is the frequency) times self- or mutual inductance is the ratio of the alternating voltage having the appropriate phase, which appears across specified terminals, to the current through the device. Commercial instruments often measure inductance from this ratio by comparing it with the voltage-to-current ratio associated with a noninductive resistor. *See* ELECTRICAL IMPEDANCE.

Precautions. Some practical precautions must be taken if accurate results are to be obtained. Any magnetic field associated with the inductor must not interact significantly with magnetic or conducting material in the vicinity of the inductor, inasmuch as the field, and therefore the inductance, would be altered. The varying magnetic field of an inductor will induce eddy currents in any nearby conducting material, which will in turn produce a magnetic field which interacts with the inductor and measuring system. Errors in a measurement of inductance may also arise from the interaction of the magnetic field of an inductor with the rest of the measuring circuit. Capacitance to other parts or to the surroundings of an inductor arising from its associated electric field will inevitably affect the impedance or apparent inductance of an inductor by a frequency-dependent amount but capacitive currents associated with screening of the measuring circuit can be routed in such a way as not to affect the measurement. *See* EDDY CURRENT; INDUCTOR.

If the magnetic circuit of an inductor includes magnetic material whose permeability depends on its previous magnetic history, or the magnetic flux caused by a direct current flowing simultaneously in the coil, its inductance will also be current- or history-dependent, and these conditions must be specified if the measurement is to be meaningful.

The electrical property of self- or mutual inductance is only defined for complete circuits. Since a measuring device or network forms part of the complete circuit when it is connected to an inductor to perform a measurement, it is necessary to ensure that either the inductance associated with the measuring circuit is negligible or that the measured quantity is defined as the change in inductance occurring when the unknown is replaced by a short circuit. The former procedure is usual for mutual inductors, and the latter for self-inductors.

Phase defect. In a mutual inductor the changing magnetic flux resulting from a varying current i in one coil induces a varying voltage V, 90° out of phase, in another coil. A perfect mutual inductor would have negligible energy losses associated with the interactions between the changing electric and magnetic fields of the device and nearby resistive conductors, including the conductors of the coils themselves. The energy loss per cycle in a practical air-cored mutual inductor amounts to only a few parts per million of the total stored magnetic energy of the device. The effect of this loss is that the phase of the voltage induced in the secondary will lead that of the current in the primary by slightly less than 90°. The departure from 90° is called the phase defect of the device. It is usual to define a configuration for a mutual inductor in which one specified end of the primary coil is connected to a specified end of the secondary coil; the effect of a phase defect can then be represented by an equivalent circuit (**Fig. 1**), in which the primary and secondary share a common fictitious path of small resistance r. This resistance has nothing to do with the actual resistance of the primary coil, which is irrelevant as far as the mutual inductance of the device is concerned.

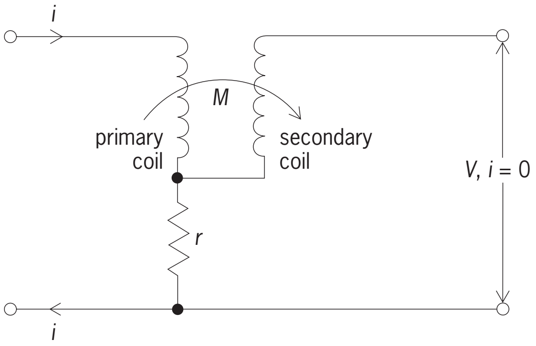


Fig. 1. Equivalent circuit representing loss in a mutual inductor.

In a self-inductor the functions of primary and secondary are combined in one coil, and therefore the voltage across the coil includes the potential drop of the current as it flows through the actual resistance of the coil. Self-inductors therefore have large, frequency-dependent phase defects, and this makes measuring them difficult. Their phase defect may be represented by an equivalent circuit in which a resistance R (approximately the actual resistance of the coil) is in series with an inductance L.

Comparing mutual inductors and deriving their values in terms of other electrical standards requires different methods from those used for self-inductors. The measurement of each will be discussed separately, taking account of the above considerations.

Measuring self-inductance. Capacitance and resistance standards have replaced inductance standards as the primary standards of impedance, and a scale of self-inductance is now established by finding the value of a self-inductor in terms of the reactance of a known capacitance C. For example, if these components are connected in parallel so that the combination appears to be purely resistive when measured in a suitable bridge at a frequency f, then Eq. (1) is satisfied.

$$2\pi f L' = \frac{1}{2\pi f C} \qquad (1)$$

See CAPACITANCE MEASUREMENT; RESISTANCE MEASUREMENT.

This procedure measures the inductance parameter associated with a parallel equivalent circuit in which a resistance R' is in parallel with the inductance of value L'. The conversion between this representation and the more usual series equivalent circuit is given by Eqs. (2), where $D = 2\pi f L'/R' = R/2\pi f L$

$$L = \frac{L'}{1+D^2} \qquad R = \frac{R'D^2}{1+D^2} \qquad (2)$$

is the reciprocal of the Q factor of the inductor.

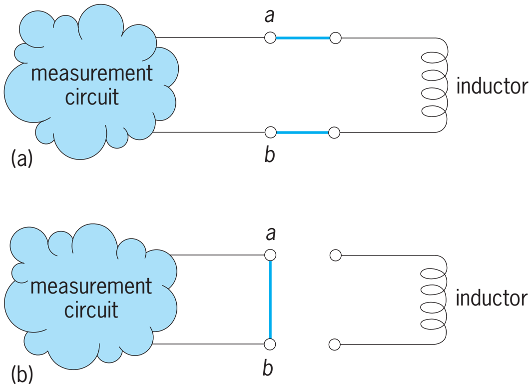


Fig. 2. Definition of the value of a self-inductor. (*a*) Circuit with self-inductor. (*b*) Circuit with self-inductor replaced by shorting link between terminals *a* and *b*

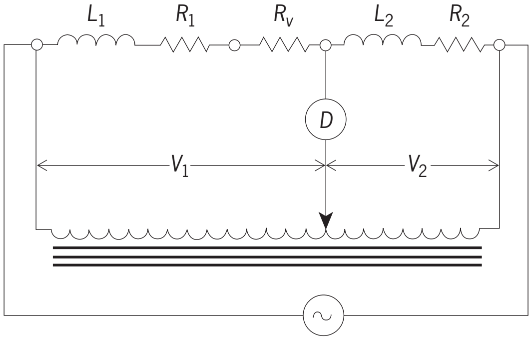


Fig. 3. Comparison of self-inductor using an inductive voltage divider.

Self-inductance should be measured as the change obtained by breaking the circuit at terminals *a* and *b* (**Fig. 2***a*) and substituting a shorting link for the inductor.

Self-inductors can be compared in a bridge in which the ratio of voltages V_1/V_2 across them is equated to the voltage ratio of a tapped autotransformer, otherwise called an inductive voltage divider (**Fig. 3**). Adjusting a noninductive resistance R_v in series with the inductor having the higher value of the ratio L/R enables the magnitude and phase of these voltage ratios to be equated when the detector is nulled. Then Eqs. (3) are satisfied.

$$L_1 = \frac{V_1}{V_2} L_2 \qquad R_1 = \frac{V_1}{V_2} R_2 - R_v \qquad (3)$$

See INDUCTIVE VOLTAGE DIVIDER.

If the inductors have any significant external magnetic field, interaction of the one with the other or with the measuring circuit can be avoided by placing the inductors far apart with their coil axes in perpendicular directions, and by constructing the measuring circuit, including the leads to the inductors, as twisted conductor pairs or as coaxial cables whose inner and outer conductors carry equal and opposite currents.

The effect on the measurement of interaction via alternating magnetic fields emanating from a measuring instrument and the leads to the inductor can be eliminated by reversing the connections to the inductor, repeating the measurement, and taking the mean of the two results. Additionally, in a bridge of the type described above which compares an unknown inductor with a standard, the effect of any coupling between the inductors can be eliminated by also reversing the connections to the standard. If the reversal of both inductors is repeated, the mean of the resulting four measurements will eliminate all unwanted magnetic couplings.

Inaccuracy of a measuring instrument might be due, among other possibilities, to insufficient rejection by the measuring instrument of the resistive component of the inductor's impedance. This can be tested by connecting a resistor of value comparable with that of the inductor and of negligible inductance in series or parallel with the inductor to be measured. Having paid due regard to whether a series or parallel equivalent circuit representing the inductor is assumed by the instrument and any necessary consequent conversion made using Eq. (2), the inductance value given by the instrument ought to be unaffected by the presence of the resistor.

Measuring mutual inductance. If losses and capacitative effects between windings and to other parts of the circuit are neglected, mutual inductance is linearly additive. That is, the voltages induced in two or more secondary coils of a mutual inductor can be added simply by connecting the windings in series. Moreover, a mutual inductor is inherently a four-terminal device having separate current and potential terminals. Its impedance is defined as the ratio of the voltage across the potential terminals to the angular frequency times the current flowing into and out of the current terminals. These two considerations make measuring and comparing mutual inductances easier and more exact than comparing self-inductances, but it is still important that no significant magnetic fields other than those of the devices being compared are produced by the measuring circuits, and that the field of one device does not interact with another.

Two mutual inductors M_1 and M_2 can be compared by using a circuit (**Fig. 4**), which again uses an inductive voltage divider as a device of known voltage ratio.

The same current can be made to flow in the primary windings of two mutual inductors by

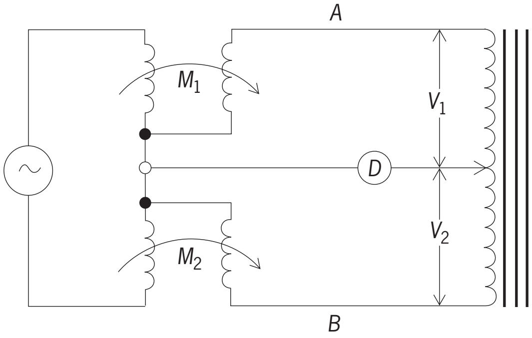


Fig. 4. Comparison of mutual inductors using an inductive voltage divider.

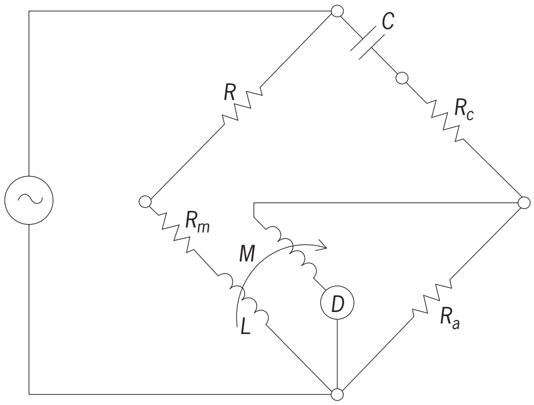


Fig. 5. Diagram of Heydweiller bridge.

connecting them in series across the bridge source. The ratio of the two voltages induced in the secondary coils can then be compared with the voltage ratio of the inductive voltage divider. Should the input impedance of the inductive voltage divider be insufficiently large so that the secondary coils are loaded by it to an unacceptable extent, it will be necessary to supply the energizing current for the inductive voltage divider directly from the source in such a way that the defining condition of negligible current in the conductors at A and B is realized. Then the ratio of the inductors is given by Eq. (4)

$$\frac{M_1}{M_2} = \frac{V_1}{V_2} \qquad (4)$$

to an accuracy of the order of $(r/2\pi fM)^2$, where r represents the loss of a mutual inductor M as in the equivalent circuit of Fig. 1. This accuracy will usually be sufficient for all ordinary purposes.

The Heydweiller or Carey-Foster bridge (**Fig. 5**) is a convenient classical four-armed bridge for mutual inductance; L and R_m are respectively the inductance and resistance of the current-carrying windings of the mutual inductor. As can be seen, the source and detector share a common terminal, which simplifies the problem of using shields to prevent stray capacitive impedances from affecting the measurement.

If the parameter r which represents the phase defect of a mutual inductor can be neglected, the bridge balance equations are Eqs. (5) and (6).

$$M = R_a(R_m + R)C \qquad (5)$$

$$L = (R_c + R_a)(R_m + R)C \qquad (6)$$

R_m must be known, or separately measured, but R can often be chosen to be much larger than R_m, so that no great accuracy for R_m is required.

A mutual inductor can also be measured by connecting the primary and secondary coils in series in the same and opposing senses, and measuring the respective total self-inductances L_1 and L_2. The mutual inductance M can then be calculated from Eq. (7).

$$M = \frac{L_1 - L_2}{4} \qquad (7)$$

Shielding of capacitive currents. In all the circuits described above, appropriate electrical screens should surround the conductors and components of the measuring circuit. These screens should be maintained at the potential of the detector terminals by connecting them to the output of a suitably adjusted potential divider connected across the source. Capacitive currents from all parts of the circuit are then intercepted and diverted back to the source so that they do not flow through the detector and influence the bridge balance.

A further improvement in technique is to construct the measurement system with fully screened components interconnected with coaxial cables. There are methods to ensure that the currents in the coaxial inner conductors are balanced by equal and opposite currents in the coaxial outer conductors. In this way unwanted magnetic interactions between the various parts of the network can also be avoided. *See* COAXIAL CABLE; ELECTRICAL MEASUREMENTS; ELECTRICAL SHIELDING; INDUCTANCE.

Bryan P. Kibble

Bibliography. B. Hague and T. R. Foord, *Alternating Current Bridge Methods*, 6th ed., 1971; B. P. Kibble and G. H. Rayner, *Coaxial A.C. Bridges*, 1984; L. Schnell (ed.), *Technology of Electrical Measurements*, 1993.

Induction coil

A device for producing a high-voltage alternating current or high-voltage pulses from a low-voltage direct current. The largest modern use of the induction coil is in the ignition system of internal combustion engines, such as automobile engines. Devices of similar construction, known as vibrators, are used as rectifiers and synchronous inverters. *See* IGNITION SYSTEM.

Figure 1 shows a typical circuit diagram for an induction coil. The primary coil, wound on the iron core, consists of only a few turns. The secondary coil, wound over the primary, consists of a large number of turns.

When the switch S is closed, the iron core becomes magnetized and therefore attracts the armature A. This automatically breaks the circuit to the coil through contact B and the armature. The armature is returned to its initial position by a spring and again makes contact with the contact B, restoring

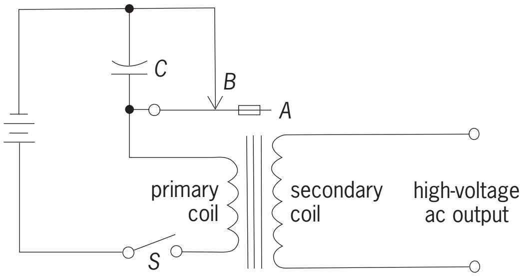


Fig. 1. Typical circuit for an induction coil.

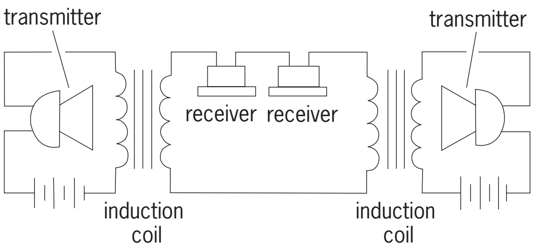


Fig. 2. Induction coils in telephone circuit.

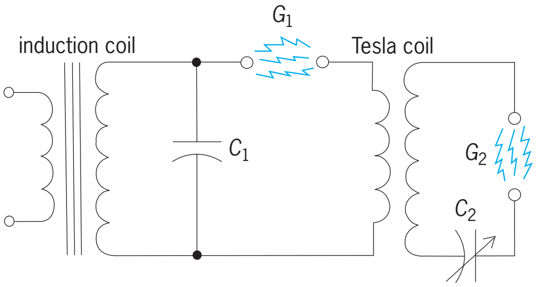


Fig. 3. Circuit diagram of Tesla coil.

the circuit to the primary coil. The cycle is then repeated rapidly.

While current is flowing in the primary coil, a magnetic field is produced. When the contact between *A* and *B* is broken, the magnetic field collapses and induces a high voltage in the secondary coil, similar to transformer action. The self-inductance of the coil must be limited; therefore, the core is a straight bundle of iron wires, which minimize eddy-current losses, rather than a closed iron circuit as is used in a transformer. *See* TRANSFORMER.

The capacitor *C* is placed across the breaker contacts to reduce the voltage across the contacts at the moment of their opening and thus reduce sparking. Sparking is caused by the induced voltage in the primary winding resulting from the collapsing magnetic field. The capacitor allows some of the energy of the magnetic field to be converted to electrostatic energy in the capacitor, rather than into heat at the contacts.

Induction coils of a different type are used in telephone circuits to step up the voltage from the transmitter and match the impedance of the line. The direct current in the circuit varies in magnitude at speech frequencies; therefore, no interrupter contacts are necessary. The battery and primary winding are connected in series with the transmitter as in **Fig. 2**. The secondary winding and the receiver are connected in series with the line. This circuitry reduces the required battery voltage.

Still another type of induction coil, called a reactor, is really a one-winding transformer designed to produce a definite voltage drop for a given current. *See* REACTOR (ELECTRICITY).

In 1892 Nicola Tesla used a form of induction coil to obtain currents of very high frequencies and high voltages. The oscillatory discharge of a Leyden jar was used as the interrupter. The terminals of the secondary of an induction coil are connected, one to the inner coating and the other to the outer coating of an insulated Leyden jar C_1 (**Fig. 3**). The circuit is completed through the primary winding of the Tesla coil, and the primary gap G_1. The coils and the Leyden jar act as a resonant circuit in the production of the high-frequency oscillation.

The primary of the Tesla coil consists of a half-dozen turns of wires wound on a nonmagnetic core. The secondary consists of many turns. The two coils are separated by air or oil as insulation. The alternation from the Leyden jar may have a frequency of several million hertz. Hence, the current induced in the secondary is not only of high voltage but also of high frequency.

Norman R. Bell

Induction heating

The heating of a nominally electrical conducting material by eddy currents induced by a varying electromagnetic field. The principle of the induction heating process is similar to that of a transformer. In **Fig. 1**, the inductor coil can be considered the primary winding of a transformer, with the workpiece as a single-turn secondary. When an alternating current flows in the primary coil, secondary currents will be induced in the workpiece. These induced currents are called eddy currents. The current flowing in the workpiece can be considered as the summation of all of the eddy currents.

In the design of conventional electrical apparatus, the losses due to induced eddy currents are minimized because they reduce the overall efficiency. However, in induction heating, their maximum effect is desired. Therefore close spacing is used between the inductor coil and the workpiece, and high-coil currents are used to obtain the maximum induced eddy currents and therefore high heating rates. *See* CORE LOSS; EDDY CURRENT.

Applications. Induction heating is widely employed in the metalworking industry for a variety of industrial processes. While carbon steel is by far the most common material heated, induction heating is also used with many other conducting materials such as various grades of stainless steel,

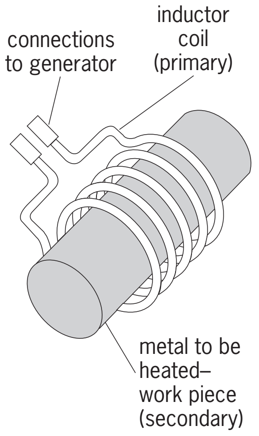


Fig. 1. Basic elements of induction heating.

aluminum, brass, copper, nickel, and titanium products. Induction heating is widely employed in (1) melting, holding, and superheating of ferrous and nonferrous metals in coreless and channel furnaces; (2) forging, forming, and rolling of slabs, billets, bars, and portions of bars; (3) heat treatment, such as hardening, annealing, and tempering applications; these processes are performed as continuous in-line systems or with automated handling of discrete parts to improve the metal's physical properties; (4) surface conditioning, such as curing of coatings, sintering, and semiconductor processing; (5) metal joining such as welding, brazing, and soldering. *See* BRAZING; HEAT TREATMENT (METALLURGY); SOLDERING; WELDING AND CUTTING OF MATERIALS.

The advantages of induction heating over the conventional processes (like fossil furnace or saltbath heating) are the following: (1) Heating is induced directly into the material. It is therefore an extremely rapid method of heating. It is not limited by the relatively slow rate of heat diffusion in conventional processes using surface-contact or radiant heating methods. (2) Because of skin effect, the heating is localized and the heated area is easily controlled by the shape and size of the inductor coil. (3) Induction heating is easily controllable, resulting in uniform high quality of the product. (4) It lends itself to automation, in-line processing, and automatic-process cycle control. (5) Startup time is short, and standby losses are low or nonexistent. (6) Working conditions are better because of the absence of noise, fumes, and radiated heat.

Heating process. The induced currents in the workpiece decrease exponentially from the surface toward the center of the material being heated. This current flowing through the material's own resistivity generates heat. The depth of heat penetration is defined as the depth to which 87% of the heat is generated. It is proportional to the square root of the material's resistivity ρ divided by the square root of its permeability μ and frequency f of the ac power applied to the inductor ($\delta_W = \sqrt{\rho/\mu f}$). The effective depth of heat penetration is therefore greater for lower frequencies and higher-resistivity materials. In magnetic materials such as steel, the depth of current penetration is less below the Curie temperature (approximately 1350°F or 730°C where the steel is magnetic) than it is above the Curie temperature. *See* CURIE TEMPERATURE; SKIN EFFECT (ELECTRICITY).

The rate at which thermal energy is required (P_{th}) can easily be calculated by the formula below.

$$\begin{aligned} P_{th} &= MC\Delta T \times 10^{-3}\ \text{kW} \\ &= 17.6M'C'\Delta T' \times 10^{-3}\ \text{kW} \end{aligned}$$

M = kilograms of product heated per second
C = specific heat, (J/kg)/°C
ΔT = temperature rise in °C
M' = pounds of product heated per minute
C' = specific heat, (Btu/lb)/°F
$\Delta T'$ = temperature rise in °F

Frequencies used in induction heating

Frequency	Power source	Application
50/60 Hz 180 Hz–3 kHz	Line Solid-state	Melting
50/60 Hz 180 Hz–10 kHz	Line Solid-state	Forging, forming, rolling
1–10 kHz 200–500 kHz	Solid-state Vacuum tube oscillators	Heat treatment and surface conditioning
10–50 kHz 200–500 kHz	Solid-state Vacuum tube oscillators	Metals joining

When the proper frequency is employed, the overall heating efficiency averages 75% for magnetic materials, and 30–50% for nonmagnetic materials. For efficient heating, the frequency employed must be high enough so that the depth of current penetration is less than one-third the diameter or cross section of the material being heated. The **table** shows the range of frequencies used for applications of the induction heating process.

Since the capital equipment cost (in dollars per kilowatt) increases with the frequency of the power supply, the lowest frequency that will properly heat the part is employed. Where high-volume production is involved, as in forging applications, line frequency and dual frequencies with separate induction coils are often employed to maximize production at minimum equipment costs. When the workpieces are small, it is necessary to use higher frequencies to efficiently heat the part. Likewise, higher frequencies must be used when it is

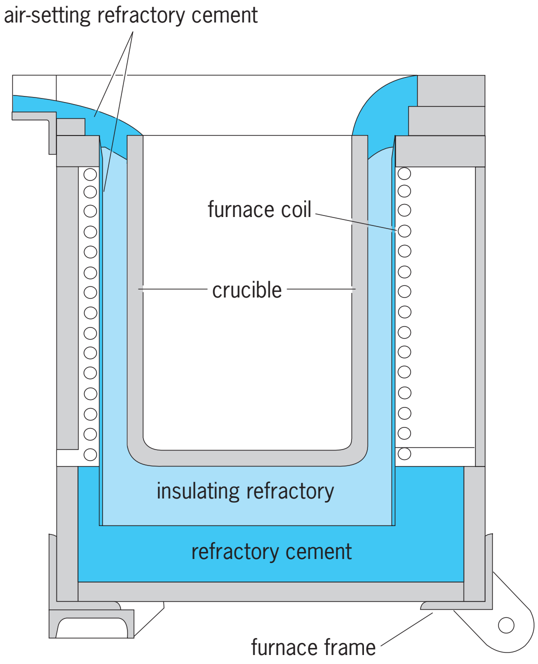


Fig. 2. Cross section of a typical induction melting furnace. (*Inductotherm Corp.*)

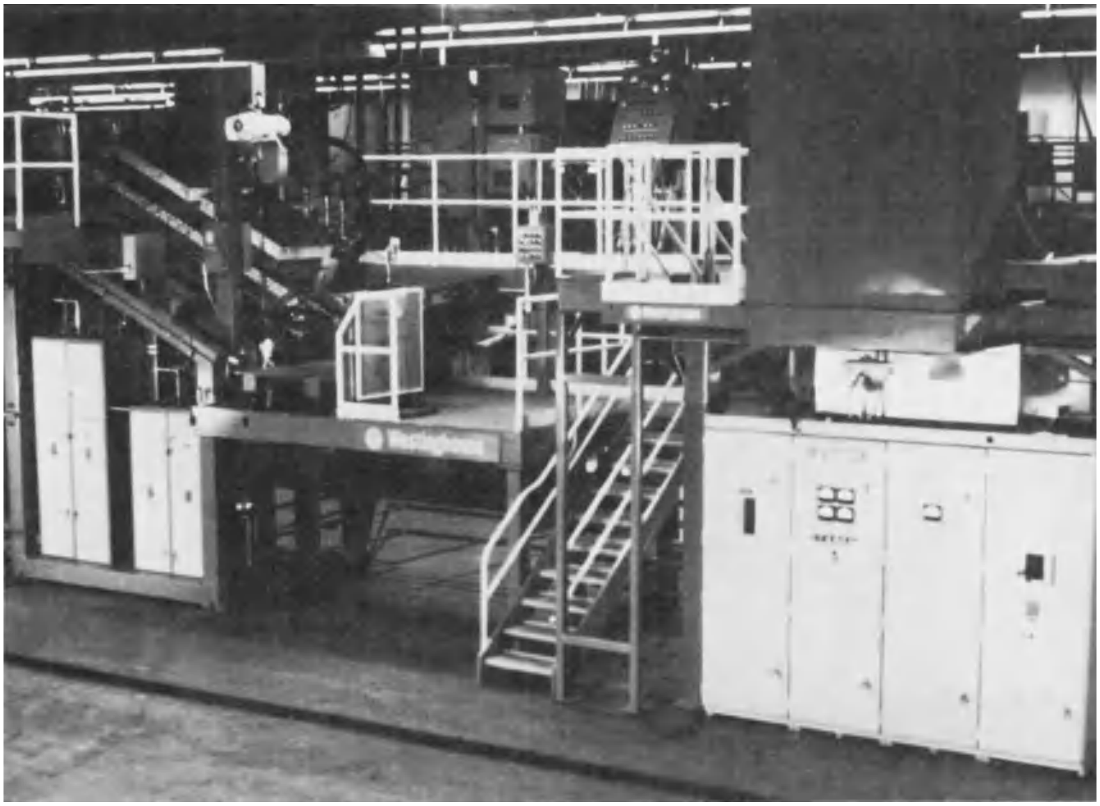


Fig. 3. A 1600-kW bar end heating system for feeding an automatic upsetter for forging flanges on automotive axles at a production rate of 720 per hour. (*Westinghouse Electric Corp.*)

necessary to concentrate the heat near the surface, as in surface hardening applications.

Power sources. The equipment used as power sources depends on the frequency required for the application. When line frequencies (generally 60 Hz) are used, suitable transformers, power factor correction capacitors, and control equipment are employed.

For higher frequencies, ranging up to 50,000 Hz, solid-state power supplies have replaced rotating equipment (motor generators). These power supplies utilize high-power thyristor solid-state devices to generate frequencies of 180 Hz to 50 kHz and ratings from 30 to 2000 kW. The conversion efficiencies of these power supplies average 90% or better as compared with the motor generator equipment with conversion efficiencies of only about 75%. *See* SEMICONDUCTOR RECTIFIER.

For frequencies over 50 kHz, power sources utilize vacuum tube oscillators to generate frequencies up to 4 MHz with power ratings from 1 to 800 kW. The conversion efficiencies of these power sources average only 60%, and therefore solid-state devices and equipment have been developed to replace them. However, the turn-off time characteristics and economics of silicon controlled rectifiers dictate the use of transistors, and therefore solid-state equipment is limited to low power ratings. *See* OSCILLATOR.

Process use. Induction heating is used for many processes as shown in the table. The construction of a typical melting furnace is shown in **Fig. 2**. A typical bar-heating system is shown in **Fig. 3**. Many different types of material-handling equipment such as scanners, indexing fixtures, conveyors, and lift and carry mechanisms are used with automated induction heating systems. *See* FURNACE CONSTRUCTION.

Induction heating differs from other methods in its high speed of heating and in its ability to generate heat from within the metal itself. In applications such as hardening, time and temperature affect the end product. Therefore slightly different heating cycles, temperatures, and materials are selected to provide the desired metallurgical properties in the end product. For other heating methods *see* ELECTRIC HEATING.

George F. Bobart

Bibliography. American Society of Metals, *Source Book on Heat Treating*, vol. 1, 1975; E. J. Davies and P. G. Simpson, *Induction Heating Handbook*, 1978; S. Zinn and S. L. Semiatin, *Elements of Induction Heating*, 1988.

Induction motor

An alternating-current motor in which the currents in the secondary winding (usually the rotor) are created solely by induction. These currents result from voltages induced in the secondary by the magnetic field of the primary winding (usually the stator). An induction motor operates slightly below synchronous speed and is sometimes called an asynchronous (meaning not synchronous) motor.

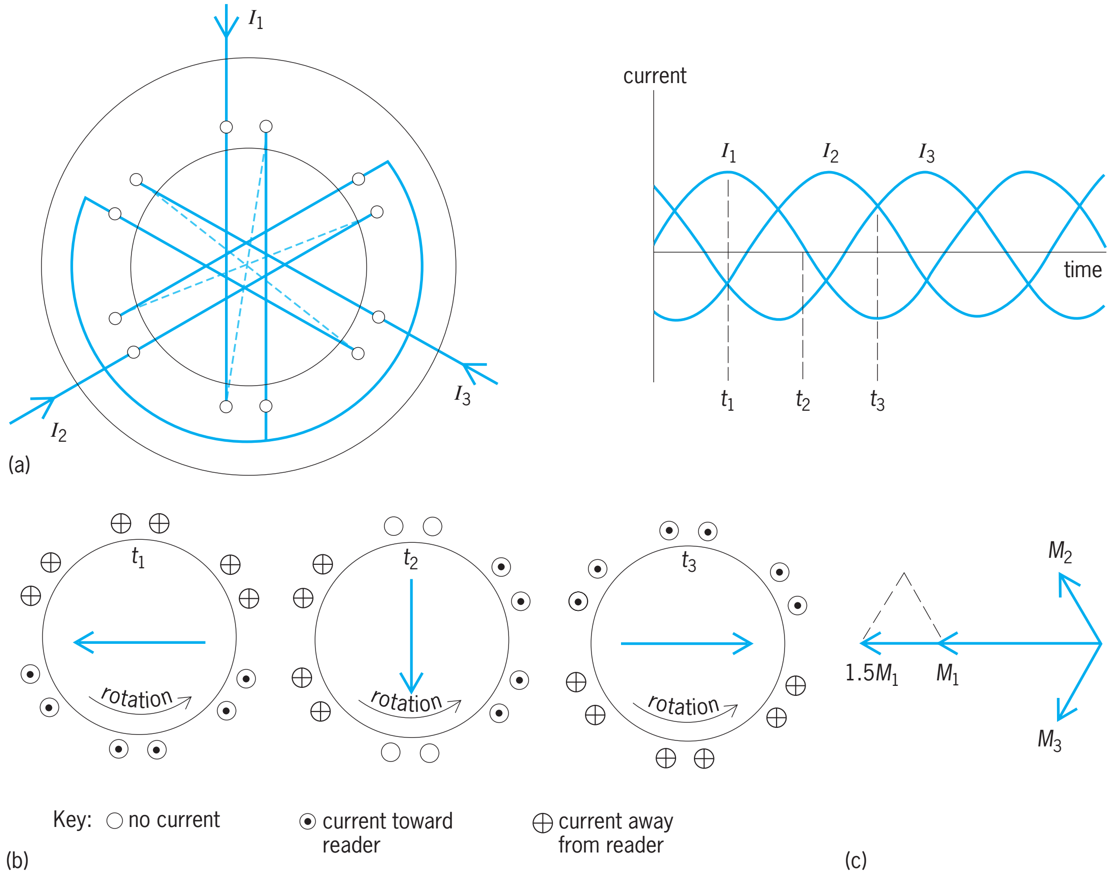


Fig. 1. Three-phase, two-pole, Y-connected stator of induction motor supplied with currents I_1, I_2, and I_3. (*a*) Stator windings and currents. (*b*) Rotating field. (*c*) Magnetomotive forces produced by stator winding.

Induction motors are the most commonly used electric motors because of their simple construction, efficiency, good speed regulation, and low cost. Polyphase induction motors come in all sizes and find wide use where polyphase power is available. Single-phase induction motors are found mainly in fractional-horsepower sizes (1 hp = 0.75 kW), and those up to 25 hp (19 kW) are used where only single-phase power is available.

Polyphase Induction Motors

There are two principal types of polyphase induction motors: squirrel-cage and wound-rotor machines. The differences in these machines is in the construction of the rotor. The stator construction is the same and is also identical to the stator of a synchronous motor. Both squirrel-cage and wound-rotor machines can be designed for two- or three-phase current.

Stator. The stator of a polyphase induction motor produces a rotating magnetic field when supplied with balanced, polyphase voltages (equal in magnitude and 90 electrical degrees apart for two-phase motors, 120 electrical degrees apart for three-phase motors). These voltages are supplied to phase windings, which are identical in all respects. The currents resulting from these voltages produce a magnetomotive force (mmf) of constant magnitude which rotates at synchronous speed. The speed is proportional to the frequency of the supply voltage and inversely proportional to the number of poles constructed on the stator.

Figure 1 is a simplified diagram of a three-phase, two-pole, Y-connected stator supplied with currents I_1, I_2, and I_3. Each stator winding produces a pulsating mmf which varies sinusoidally with time. The resultant mmf of the three windings (Fig. 1*c*) is constant in magnitude and rotates at synchronous speed. Figure 1*b* shows the direction of the mmf in the stator for times t_1, t_2, and t_3 shown in Fig. 1*a* and shows how the resultant mmf rotates. The synchronous speed N_s is shown by Eq. (1), where f is the frequency

$$N_s = \frac{120f}{p}\,\text{rpm} \tag{1}$$

in hertz and p is the number of stator poles. For any given frequency of operation, the synchronous speed is determined by the number of poles. For 60-Hz frequency, a two-pole motor has a synchronous speed of 3600 rpm; a four-pole motor, 1800 rpm; and so on. For details of stator windings *see* WINDINGS IN ELECTRIC MACHINERY

Squirrel-cage rotor. **Figure 2** shows the bars, end rings, and cooling fins of a squirrel-cage rotor. The bars are skewed or angled to prevent cogging (operating below synchronous speed) and to reduce noise. The end rings provide paths for currents that result from the voltages induced in the rotor bars by the stator flux. The number of poles on a squirrel-cage

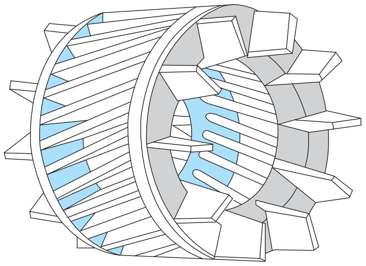


Fig. 2. Bars, end rings, and cooling fins of a squirrel-cage rotor.

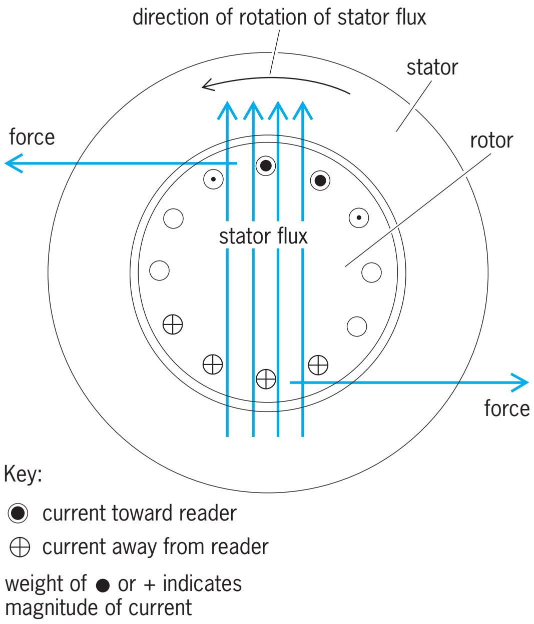


Fig. 3. Forces on the rotor winding.

rotor is always equal to the number of poles created by the stator winding.

Figure 3 shows how the two motor elements interact. A counterclockwise rotation of the stator flux causes voltages to be induced in the top bars of the rotor in an inward direction. These currents interact with the stator flux and produce a force on the rotor bars in the direction of the rotation of the stator flux.

When not driving a load, the rotor approaches synchronous speed N_s. At this speed there is no motion of the flux with respect to the rotor conductors. As a result, there is no voltage induced in the rotor and no rotor current flows. As load is applied, the rotor speed decreases slightly, causing an increase in rotor voltage and rotor current and a consequent increase in torque developed by the rotor. The reduction in speed is therefore sufficient to develop a torque equal and opposite to that of the load. Light loads require only slight reductions in speed; heavy loads require greater reduction. The difference between the synchronous speed N_s and the operating speed N is the slip speed. Slip s is conveniently expressed as a percentage of synchronous speed, as in Eq. (2).

$$s = \frac{N_s - N}{Ns} \times 100\% \qquad (2)$$

When the rotor is stationary, a large voltage is induced in the rotor. The frequency of this rotor voltage is the same as that of the supply voltage. The frequency f_2 of rotor voltage at any speed is shown by Eq. (3), where f_1 is the frequency of the supply

$$f_2 = f_1 s \qquad (3)$$

voltage and s is the slip expressed as a decimal. The voltage e_2 induced in the rotor at any speed is shown by Eq. (4), where e_{2s} is the rotor voltage at standstill.

$$e_2 = (e_{2s})s \qquad (4)$$

The reactance x_2 of the rotor is a function of its standstill reactance x_{2s} and slip, as shown by Eq. (5). The

$$x_2 = (x_{2s})s \qquad (5)$$

impedance of the rotor at any speed is determined by the reactance x_2 and the rotor resistance r_2. The rotor current i_2 is shown by Eq. (6). In the equation, for

$$\begin{aligned} i_2 &= \frac{e_2}{\sqrt{r_2^2 + x_2^2}} \\ &= \frac{(e_{2s})s}{\sqrt{r_2^2 + (x_{2s})^2 s^2}} = \frac{e_{2s}}{\sqrt{\left(\frac{r_2}{s}\right)^2 + (x_{2s})^2}} \end{aligned} \qquad (6)$$

small values of slip, the rotor current is small and possesses a high power factor. When slip becomes large, the r_2/s term becomes small, current increases, and the current lags the voltage by a large phase angle. Standstill (or starting) current is large and lags the voltage by 50–70°. Only in-phase, or unity-power-factor, rotor currents are in space phase with the air-gap flux and can therefore produce torque. The current i_2 contains both a unity power-factor component i_p and a reactive component i_r. The maximum value of i_p and therefore maximum torque are obtained when slip is of the correct value to make r_2/s equal to x_{2s}. If the value of r_2 is changed, the slip at which maximum torque is developed must also change. If r_2 is doubled and s is doubled, the current i_2 is not changed and the torque is unchanged.

This feature provides a means of changing the speed-torque characteristics of the motor. In **Fig. 4**, curve 1 shows a typical characteristic curve of an

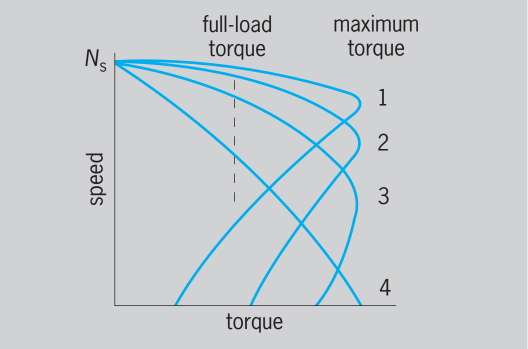


Fig. 4. Speed-torque characteristic of polyphase induction motor.

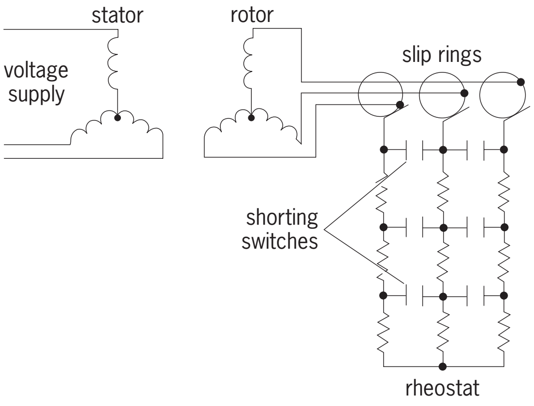


Fig. 5. Connections of wound-rotor induction motor.

induction motor. If the resistance of the rotor bars were doubled without making any other changes in the motor, it would develop the characteristic of curve 2, which shows twice the slip of curve 1 for any given torque. Further increases in the rotor resistance could result in curve 3. When r_2 is made equal to x_{2s}, maximum torque will be developed at standstill as in curve 4. These curves show that higher resistance rotors give higher starting torque. However, since the motor's normal operating range is on the upper portion of the curve, the curves also show that a higher-resistance rotor results in more variation in speed from no load to full load (or poorer speed regulation) than the low-resistance rotor. Higher-resistance rotors also reduce motor efficiency. Except for their characteristic low starting torque, low-resistance rotors would be desirable for most applications.

Wound rotor. A wound-rotor induction motor can provide both high starting torque and good speed regulation. This is accomplished by adding external resistance to the rotor circuit during starting and removing the resistance after speed is attained.

The wound rotor has a polyphase winding similar to the stator winding and must be wound for the same number of poles. Voltages are induced in these windings just as they are in the squirrel-cage rotor bars. The windings are connected to slip rings so that connections may be made to external impedances, usually resistors, to limit starting currents, improve power factor, or control speed.

Figure 5 shows the connection of a rheostat used to bring a wound-rotor motor up to speed. The rheostat limits the starting current drawn from the supply to a value less than that required by a squirrel-cage motor. The resistance is gradually reduced to bring the motor up to speed. By leaving various portions of the starting resistances in the circuit, some degree of speed control can be obtained, as in Fig. 4. However, this method of speed control is inherently inefficient and converts the motor into a variable-speed motor, rather than an essentially constant-speed motor. For other means of controlling speed of polyphase induction motors and for other types of ac motors *see* ALTERNATING-CURRENT MOTOR

Single-Phase Induction Motors

Single-phase induction motors display poorer operating characteristics than polyphase machines, but are used where polyphase voltages are not available. They are most common in small sizes ($^1/_2$ hp or 0.4 kW or less) in domestic and industrial applications. Their particular disadvantages are low power factor, low efficiency, and the need for special starting devices.

The rotor of a single-phase induction motor is of the squirrel-cage type. The stator has a main winding which produces a pulsating field. At standstill, the pulsating field cannot produce rotor currents that will act on the air-gap flux to produce rotor torque. However, once the rotor is turning, it produces a cross flux at right angles in both space and time with the main field and thereby produces a rotating field comparable to that produced by the stator of a two-phase motor.

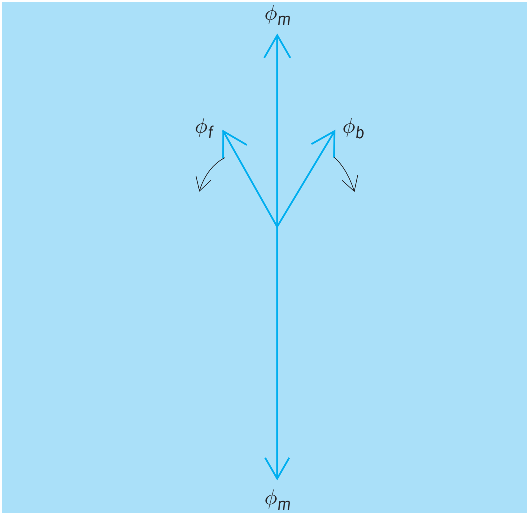


Fig. 6. Fluxes associated with the single-phase induction motor.

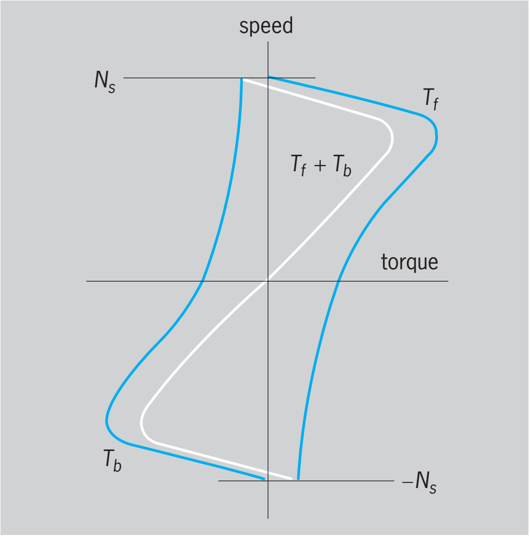


Fig. 7. Torques produced in the single-phase induction motor.

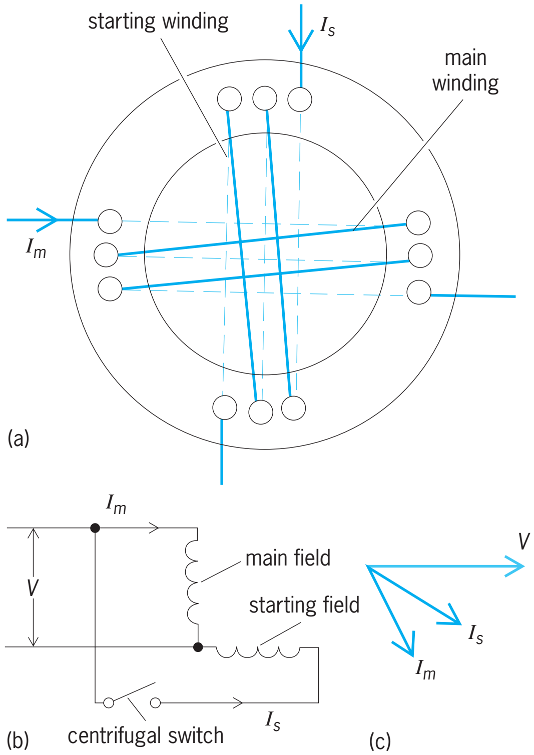


Fig. 8. Split-phase motor. (*a*) Windings. (*b*) Winding connections. (*c*) Vector diagram.

An explanation of this is based on the concept that a pulsating field is the equivalent of two oppositely rotating fields of one-half the magnitude of the resultant pulsating field. In **Fig. 6**, ϕ_m is the maximum value of the stator flux ϕ, which is shown only by its two components ϕ_f and ϕ_b, which represent the two oppositely rotating fields of constant equal magnitudes of $\phi_m/2$. Each component ϕ_f and ϕ_b produces a torque T_f and T_b on the rotor. **Figure 7** shows that the sum of these torques is zero when speed is zero. However, if started, the sum of the torques is not zero and rotation will be maintained by the resultant torque.

This machine has good performance at high speed. However, to make this motor useful, it must have some way of producing a starting torque. The method by which this starting torque is obtained designates the type of the single-phase induction motor.

Split-phase motor. This motor has two stator windings, the customary main winding and a starting winding located 90 electrical degrees from the main winding, as in **Fig. 8***a*. The starting winding has fewer turns of smaller wire, to give a higher resistance-to-reactance ratio, than the main winding. Therefore their currents I_m (main winding) and I_s (starting winding) are out of time phase, as in Fig. 8*c*, when the windings are supplied by a common voltage V. These currents produce an elliptical field (equivalent to a uniform rotating field superimposed on a pulsating field) which causes a unidirectional torque at standstill. This torque will start the motor. When sufficient speed has been attained, the circuit of the starting winding can be opened by a centrifugal switch and the motor will operate with a characteristic illustrated by the broken-line curve of Fig. 7.

Capacitor motor. The stator windings of this motor are similar to the split-phase motor. However, the starting winding is connected to the supply through a capacitor (**Fig. 9***a*). This results in a starting winding current which leads the applied voltage. The motor then has winding currents at standstill which are nearly 90° apart in time, as well as 90° apart in space. High starting torque and high power factor are therefore obtained. The starting winding circuit can be opened by a centrifugal switch when the motor comes up to speed. A typical characteristic is shown in Fig. 9*c*.

In some motors two capacitors are used. When the motor is first connected to the voltage supply, the two capacitors are used in parallel in the starting circuit. At higher speed one capacitor is removed by a centrifugal switch, leaving the other in series with the starting winding. This motor has high starting torque and good power factor.

Shaded-pole motor. This motor is used extensively where large power and large starting torque are not required, as in fans. A squirrel-cage rotor is used with a salient-pole stator excited by the ac supply. Each salient pole is slotted so that a portion of the pole face can be encircled by a short-circuited winding, or shading coil.

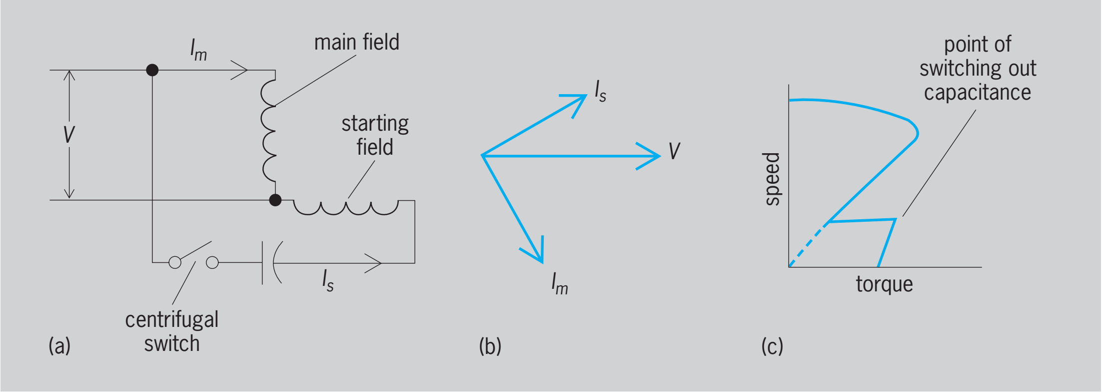


Fig. 9. Capacitor motor. (*a*) Winding connections. (*b*) Vector diagram. (*c*) Characteristic.

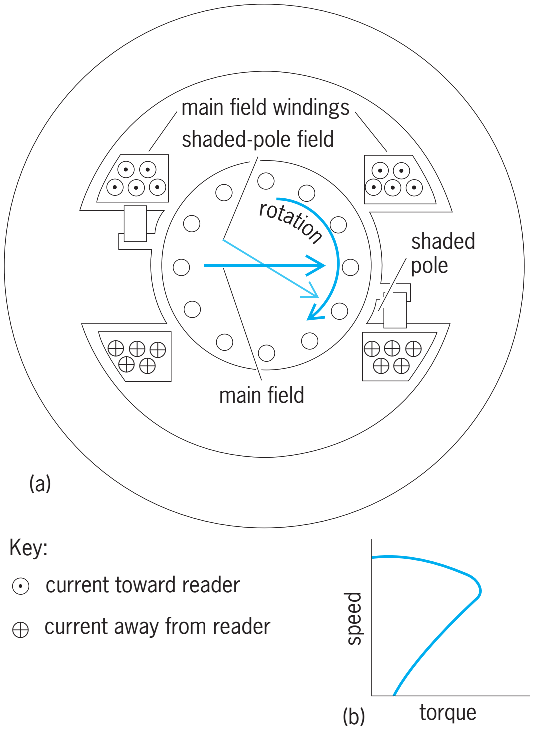


Fig. 10. Shaded-pole motor. (*a*) Cross-sectional view. (*b*) Typical characteristic.

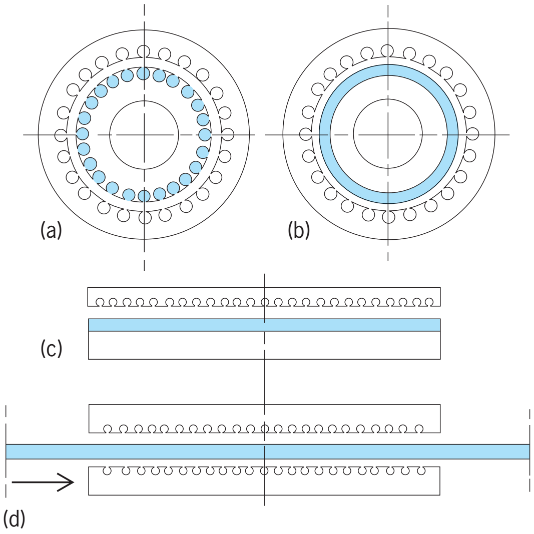


Fig. 11. Evolution of linear induction motor. (*a*) Polyphase squirrel-cage induction motor. (*b*) Conduction sheet motor. (*c*) One-sided long-secondary linear motor with short-flux return yoke. (*d*) Double-sided long-secondary linear motor.

The main winding produces a field between the poles as in **Fig. 10**. The shading coils act to delay the flux passing through them, so that it lags the flux in the unshaded portions. This gives a sweeping magnetic action across the pole face, and consequently across the rotor bars opposite the pole face, and results in a torque on the rotor. This torque is much smaller than the torque of a split-phase motor, but it is adequate for many operations. A typical characteristic of the motor is shown in Fig. 10*b*. *See* HYSTERESIS MOTOR; RELUCTANCE MOTOR; REPULSION MOTOR.

Linear motor. **Figure 11** illustrates the arrangements of the elements of the polyphase squirrel-cage induction motor. The squirrel cage (secondary) is embedded in the rotor in a manner to provide a close magnetic coupling with the stator winding (primary). This arrangement provides a small air gap between the stator and the rotor. If the squirrel cage is replaced by a conducting sheet as in Fig. 11*b*, motor action can be obtained. This machine, though inferior to that of Fig. 11*a*, will function as a motor. If the stator windings and iron are unrolled (rectangular laminations instead of circular laminations), the arrangement of the elements will take a form shown in Fig. 11*c*, and the field produced by polyphase excitation of the primary winding will travel in a linear direction instead of a circular direction. This field will produce a force on the conducting sheet that is in the plane of the sheet and at right angles with the stator conductors. A reversal of the phase rotation of the primary voltages will reverse the direction of motion of the air-gap flux and thereby reverse the force on the secondary sheet. No load on the motor corresponds to the condition when the secondary sheet is moving at the same speed as the field produced by the primary. For an arrangement of Fig. 11*c*, there is a magnetic attraction between the iron of the stator and the iron of the secondary sheet. For some applications, this can be a serious disadvantage of the one-sided motor of Fig. 11*c*. This disadvantage can be eliminated by use of the double-sided arrangement of Fig. 11*d*. Here the primary iron of the upper and lower sides is held together rigidly, and the forces that are normal to the plane of the sheet do not act on the sheet.

Applications. In conventional transportation systems, traction effort is dependent on the contact of the wheels with the ground. In some cases, locomotives must be provided with heavy weights to keep the wheels from sliding when under heavy loads. This disadvantage can be eliminated by use of the linear motor. With the linear motor with air cushions, friction loss can be reduced and skidding can be eliminated. **Figure 12** illustrates the application of the double-sided linear motor for high-speed transportation.

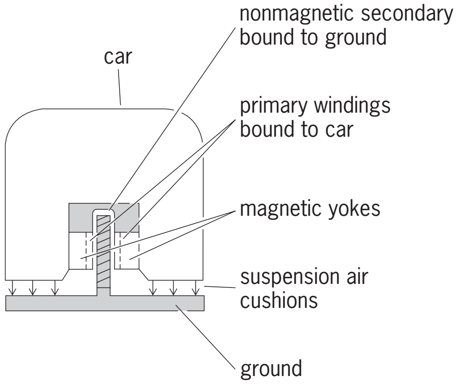


Fig. 12. Drawing of linear induction-motor configuration. Magnetic interaction supplies the forces between the car and the ground. (*Le Moteur Linéaire and Société de l'Aérotrain*)

Conveying systems that are operated in limited space have been driven with linear motors. Some of these, ranging from $^{1}/_{2}$ to 1 mi (0.8 to 1.6 km) in length, have worked successfully.

Because of the large effective air gap of the linear motor, its magnetizing current is larger than that of the conventional motor. Its efficiency is somewhat lower, and its cost is high. The linear motor is largely in the experimental stage.

Albert G. Conrad

Bibliography. P. L. Alger, *Induction Machines: Their Behavior and Use*, 2d ed., 1970, revised 1995; S. J. Chapman, *Electric Machinery Fundamentals*, 3d ed., 1998; P. L. Cochran, *Polyphase Induction Motors*, 1989; A. E. Fitzgerald, C. Kingsley and S. D. Umans, *Electric Machinery*, 5th ed., 1990; J. F. Gieras, *Linear Induction Drives*, 1994; M. Poloujadoff, *The Theory of Linear Induction Machinery*, 1980.

Inductive coordination

The avoidance of inductive interference. Electric power systems, like almost everything run by electricity, depend on internal electric and magnetic fields; some of these fields find their way into the environment. The strongest of these fields can then induce voltages and currents in nearby devices and equipment and, in some cases, can interfere with the internal fields being used by electrical equipment in the vicinity. These induced voltages and currents, which are due to the coupling between the energized source and the electrical equipment, are called inductive interference. *See* ELECTRIC FIELD; ELECTROMAGNETIC INDUCTION.

In the process of the generation and transmission of large amounts of power, power systems necessarily utilize high currents and high voltages and, further, use overhead power lines that must be in proximity to the equipment they supply. This situation can cause potential inductive coordination problems. Avoiding such problems puts constraints on the design, operation, maintenance, and siting of the various electric power system components. *See* TRANSMISSION LINES.

Any current-carrying conductor (or device) generates electric and magnetic fields, which may radiate to the environment and interfere with nearby devices. Total shielding is sometimes possible, as with underground or overhead cables. In most cases, however, shielded cables are not practical and air-insulated overhead power lines are used. They create electric and magnetic fields that couple with objects in the immediate area, producing induced voltages and currents.

Figure 1 shows the most common example of electric and magnetic induction. Objects coupled with the power lines include the power lines themselves, telephone lines, fences, tracks, and other objects within the power-line right-of-way or corridor space. Coupling is stronger near the source. Induced voltages and currents are affected by many design parameters, such as conductor size, phase spacing, geometry line configuration, object size, distance from the source, line load currents, and operating conditions. Certain induced voltages and currents could reach unsafe values.

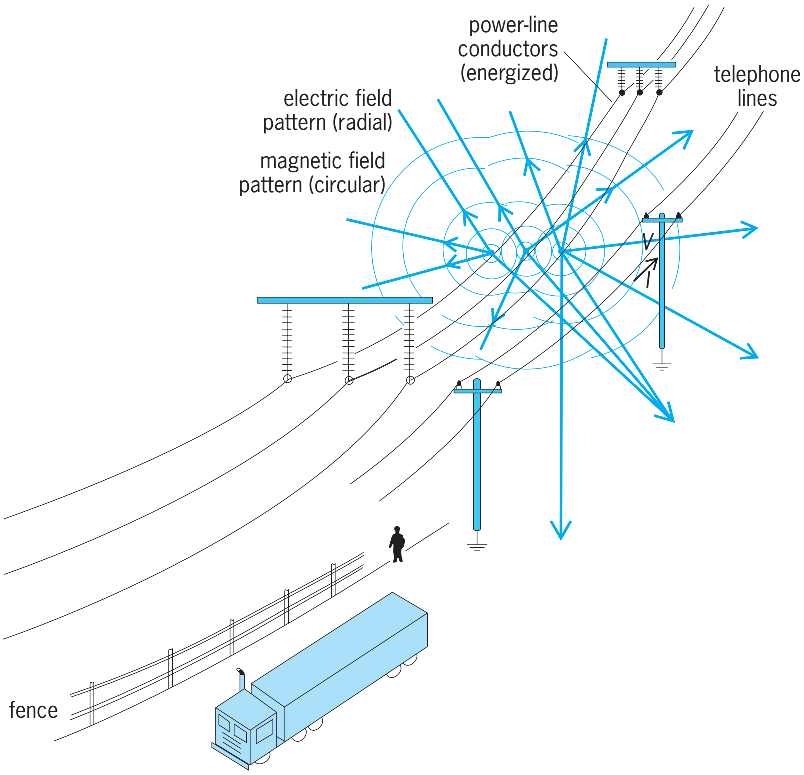


Fig. 1. Power-line field pattern and inductive coupling. Coupling produces induced voltages (V) and currents (I).

Design considerations. Inductive coupling can be reduced by proper configuration, circuit design, and siting. Generating stations and large switching stations built on remote sites and well shielded by enclosures do not produce significant stray fields. Well-shielded underground substations, manholes, cables, and distribution equipment also do not cause interference. Overhead power lines are, however, air-insulated, not shielded, and must of course be built in inhabited areas. These overhead lines cause practically all of the problems due to inductive coupling. For this reason and for safety considerations, power lines are restricted as far as possible to specific corridors or rights of way. Spacing between them and the requirements of their surroundings are considered and carefully calculated to minimize possible interference. These corridors are often shared by telephone lines, communication circuits, railroads, and sometimes trolley buses, each of which must be considered for possible inductive coupling.

Power-line inductive coordination. Modern telephone and communication circuits are well shielded and rarely encounter interference from nearby power lines. In some communication systems extensive transposition is used to reduce such interference. However, where a long parallel exposure exists, inductive coupling can be reduced by balancing the operation of the power line and by transposition of power and communication lines (**Fig. 2**).

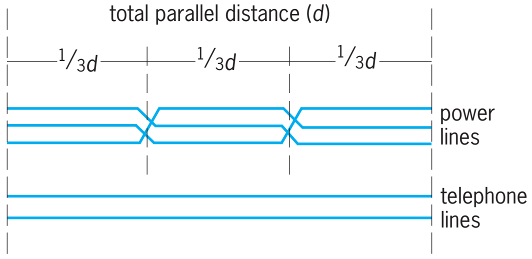


Fig. 2. Transposition: the induced voltages and currents are canceled by transposing the power line as shown.

Transposition of power lines reduces the electrical imbalances with significant reduction in induced voltages and currents in the exposed parallel telephone circuits. Fences, long irrigation pipes, and large ungrounded objects within the right of way may experience considerable inductive coupling and must be grounded for safety. *See* GROUNDING.

Special cases. In some cases, such as electric trains or unbalanced operating conditions, the ground must be used as a partial return conductor and transposition is impossible. The electric loop in such cases is composed of the wires supplying the train, the rail, and the earth. Other cases, such as solid-state devices in power conditioners, motor controllers for energy conservation, or variable-speed applications, have also created inductive coupling problems. In these cases, it would be desirable to isolate the power circuits geographically. This is usually not feasible, however, and other measures must be found, such as properly distributed capacitor banks within the power system, addition of specially designed filters, or employment of specifically designed, low-harmonic-content devices. *See* ELECTRICAL INTERFERENCE; ELECTRICAL SHIELDING.

Aly A. Mahmoud

Bibliography. Electric Power Research Institute, *Transmission Line Reference Book—345 KV And above*, 2d ed., 1982; D. G. Fink and H. W. Beaty (eds.), *Standard Handbook for Electrical Engineers*, 14th ed., 2000; Institute of Electrical and Electronics Engineers, *IEEE Recommended Practice for Inductive Coordination of Electric Supply and Communication Lines*, rev. ed., 1993.

Inductive voltage divider

An autotransformer that has its winding divided into a number of equal-turn sections (usually 10) so that when an alternating voltage V is applied to the whole winding, the voltage across each section is nominally $V/10$. The progressive voltages from one end to the section junctions are thus $V/10$, $2V/10$, $3V/10$, . . . , and $9V/10$. This voltage division can be realized with errors considerably less than 1 part per million of V, and such units therefore find wide use as standards of ac voltage ratio in the discipline of electrical measurements. *See* TRANSFORMER.

The division of voltage will be in error if there are differences of resistance and leakage inductance from section to section, and these errors will be significant if the differences are significant in relation to the input impedance of the winding. Leakage inductance is caused by that very small fraction of the flux from one section's winding which fails to thread the rest of the windings. The most commonly used constructional technique for minimizing such errors is to take 10 equal lengths of insulated copper wire and twist them into a "rope." The rope is wound onto a toroidal core made of thin, high-permeability, low-iron-loss magnetic material. The strands of the winding are then connected in series so that each strand forms the winding of one section of the 10-section divider (**Fig. 1**). The resistances of the sections are very nearly equal since the strands are the same length, and the leakage inductances are also closely equal and small because of the close flux coupling of this type of winding. The low-reluctance magnetic path of the core ensures a very high value of input impedance. Thus, voltage division at low audio frequencies can be accurate to a few parts in 10^8 of V.

If an impedance is connected across one section of an inductive voltage divider, currents will flow (**Fig. 2**) so as to produce near-zero ampere-turns on the core. Because of the close coupling of the sectional windings, only a very small proportion of the total opposing fluxes fails to thread the other sections. Hence the ratio of output to input voltage of an inductive voltage divider with one section loaded is only slightly affected by the very small leakage inductance (typically a few microhenrys in a rope-type winding) plus the ohmic resistance of the wire (typically a few tens of milliohms). Thus a single-decade

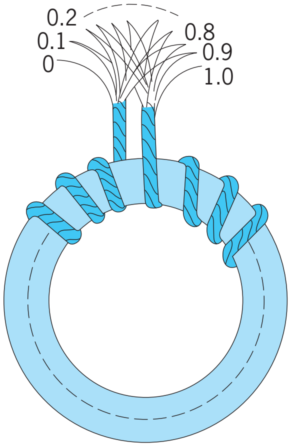


Fig. 1. Rope winding on toroidal core making single-decade inductive voltage divider.

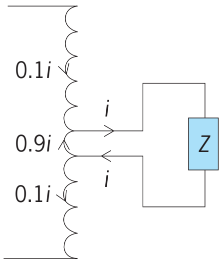


Fig. 2. Inductive voltage divider with an impedance *Z* across one section.

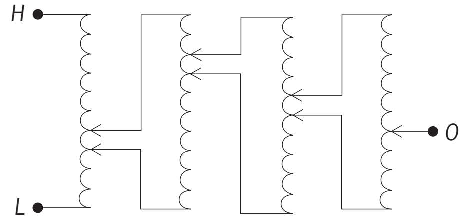


Fig. 3. **Four-decade inductive voltage divider. Input is voltage between *H* and *L*; output is voltage between *O* and *L*.**

inductive voltage divider has the characteristic of high-input impedance with low-output impedance, and because of this it is possible to connect units in tandem (**Fig. 3**) to provide a multidecade system, giving very fine resolution of the voltage division.

A common commercial application of inductive voltage dividers is as ratio arms in impedance bridges where one impedance is a transducer of a physical property (for example, temperature, force, or displacement) which is compared with a fixed reference impedance by using an inductive voltage divider as the variable ratio device. *See* INDUCTANCE MEASUREMENT.

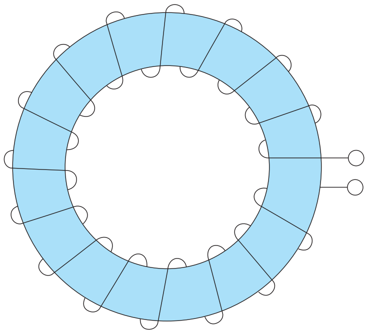
Inductive voltage dividers operate most accurately in the frequency range 20–1592 Hz but can be constructed to operate at frequencies up to 1 MHz. They are usually designed to accept an input voltage of up to about 0.25 times the frequency in hertz. *See* VOLTAGE MEASUREMENT. T. A. Deacon; Bryan P. Kibble

Bibliography. T. A. Deacon and J. J. Hill, Two-stage inductive voltage dividers, *Proc. Inst. Elec. Eng.* (*UK*), 115(6):888–892, June 1968; J. J. Hill and T. A. Deacon, Theory, design and measurement of inductive voltage dividers, *Proc. Inst. Elec. Eng.* (*UK*), 115(5):727–735, May 1968.

Inductor

A device for introducing inductance into a circuit. The term covers devices with a wide range of uses, sizes, and types, including components for electric-wave filters, tuned circuits, electrical measuring circuits, and energy storage devices.

Inductors are classified as fixed, adjustable, and variable. All are made either with or without magnetic cores. Inductors without magnetic cores are called air-core coils, although the actual core material may be a ceramic, a plastic, or some other nonmagnetic material. Inductors with magnetic cores are called iron-core coils. A wide variety of magnetic materials are used, and some of these contain very little iron. Magnetic cores for inductors for low-frequency, or high-energy storage, use are most commonly made from laminations of silicon steel. Some iron-core inductors with cores of compressed powdered iron, powdered permalloy, or ferrite are more suitable for higher-frequency applications.



Toroidal coil, a type of fixed inductor.

Fixed inductors. In fixed inductors coils are wound so that the turns remain fixed in position with respect to each other. If an iron core is used, any air gap it has is also fixed and the position of the core remains unchanged within the coil.

A toroidal coil is a fixed inductor wound uniformly around a toroidal form (see **illus.**). Because of the closed magnetic circuit, such an inductor has practically no leakage flux and is little affected by the presence of stray magnetic fields. High-accuracy standard inductors are commonly made in this form. Powdered cores are used to increase the Q of the coil and reduce the size required for a specified inductance. Ceramic-core toroidal coils supported in cork are used as standard inductors of high stability and accuracy.

Adjustable inductors. These either have taps for changing the number of turns desired, or consist of several fixed inductors which may be switched into various series or parallel combinations.

Variable inductors. Such inductors are constructed so that the effective inductance can be changed. Means for doing this include (1) changing the permeability of a magnetic core; (2) moving the magnetic core, or part of it, with respect to the coil or the remainder of the core; and (3) moving one or more coils of the inductor with respect to one or more of the other coils, thereby changing mutual inductance. *See* ALTERNATING-CURRENT CIRCUIT THEORY; INDUCTANCE. Burtis L. Robertson; Wilson S. Pritchett

Industrial ecology

The multidisciplinary study of industrial and economic systems and their linkages with fundamental natural systems. Industrial ecology incorporates research involving energy supplies, materials, technologies, and technological systems; physical, biological, and social sciences; economics; law; and business management. Industrial and economic systems are viewed not in isolation but in their cultural and ecological context. Both demand-side (consumer) and supply-side (producer) activities are included, as are all sectors of economic activity, such as mining, agriculture, forestry, fisheries, manufacturing, and service activities. Industrial ecology also includes subsistence activities at the fringes of formal economic systems, which generate a number of important impacts on natural systems. Industrial ecology provides the understanding to support the reasoned improvement of the economic, environmental, and social efficiency of current industrial systems.

History. There is an obvious appeal to thinking of industrial systems as analogous to natural ecological systems, especially when considering manufacturing activities and their associated environmental impacts. Ecological systems are characterized by a high degree of cycling of materials: the wastes from one organism become food for another. Until recently, however, most manufacturing activities were more linear: materials were used in processes and products only once, then became wastes with associated environmental impacts. Thus, the early development of industrial ecology reflected a desire to apply the cycling model characteristic of biological communities to manufacturing, with the residual materials from one process or product feeding into another. It was in this spirit that a few uses of the term "industrial ecology" began to occur in the 1970s and 1980s. In Japan, for example, the Ministry of Industry and International Trade explored such a concept in the early 1970s to suggest a model for structuring the country's industrial system.

In the late 1980s and early 1990s, industrial ecology began to evolve from an intuitively appealing metaphor to an actual field of study. Among the reasons was the recognition of the growing inability of existing environmental policy, focused on end-of-pipe emissions and remediation, and reductionist environmental science research, to address complex, systemic environmental perturbations such as global climate change, loss of biodiversity and habitat, and degradation of air, water, and soil resources. Moreover, private firms were being affected by rapidly increasing costs of environmental compliance and associated liabilities, combined with increasing public and customer pressure for enhanced environmental performance. Traditional compliance activities were not sufficient to respond to these pressures, especially in sectors such as heavy manufacturing and chemical production, which had significant environmental issues. The fundamental problem with the traditional approach is that it treated environmental issues as "overhead," that is, as ancillary to the principal activities of an individual, a firm, or, indeed, society as a whole. Industrial ecology became the framework within which environmental issues began to be treated as strategic, that is, as integral to the primary activities of firms and society.

Given these incentives, significant progress in both the theory of industrial ecology and its application through specific tools and methodologies soon followed. Among the methodologies which have become more sophisticated and accepted over this time are life-cycle assessment (LCA), developed initially in rigorous form by the Society of Environmental Toxicology and Chemistry. Life-cycle assessment is an increasingly standardized methodology which identifies the environmental impacts associated with the life cycle of a material in a specific application or, in some cases, a product, thus identifying opportunities for improvement in environmental performance. This improvement may be accomplished in a number of ways, including reductions of waste through better engineering of manufacturing processes, or substitution of less toxic materials in processes and products. Another industrial ecology application is through design for environment (DFE), developed initially by the United States electronics sector. Design for environment is a module of a broader concurrent engineering approach, design for X (DFX), where X is a desirable product characteristic such as manufacturability, safety, testability, or simplicity of materials and parts. Design for X methodologies are often used in the design and manufacture of complex articles such as computers or consumer electronics, so developing a design for environment approach made adoption of environmentally conscious manufacturing and product design easier in those manufacturing sectors. In general, life-cycle assessment methodologies are useful in evaluating a particular material in a specific application, while design for environment tends to emphasize the importance of artifact design choices and technologies, and is used more in sectors manufacturing complex articles such as electronic devices.

Industrial ecology is still a nascent field. While there are some textbooks and courses in industrial ecology, primarily in engineering schools, it still somewhat reflects its origins in manufacturing organizations. Thus, significant foundational work remains to be done to understand the industrial ecology of service economies in developed countries, on the one hand, and its application to subsistence economic activity in developing countries, on the other. Biomass sectors, such as forestry, agriculture, and fisheries, are increasingly beginning to adopt industrial ecology approaches, but in general lag behind the manufacturing sectors in doing so.

Conceptual framework. Industrial ecology can best be understood by placing it in a framework illustrating its relationships to concepts such as sustainable development, and methodologies such as design for environment and life-cycle assessment (**Fig. 1**). The highest level of the framework is the concept of sustainability, frequently framed in terms of sustainable development, or development that meets the needs of the current generation without compromising the ability of future generations to meet theirs. While many people find this a desirable concept, it is important to recognize that a number of potentially sustainable worlds can be defined, some more desirable than others. Thus, while a desirable sustainable state is important to provide a general goal toward which environmentally preferable activities can be directed, it cannot yet be defined with any specificity, or assumed as an outcome, especially in the absence of continued work in industrial ecology and related fields.

Evolution along at least three dimensions—the ethical dimension, the institutional dimension, and the science and technology dimension—is required if a desirable sustainable state is to be defined and achieved. Industrial ecology focuses on scientific and technological, not ethical or institutional, issues; it supports the evolution of the science and technology

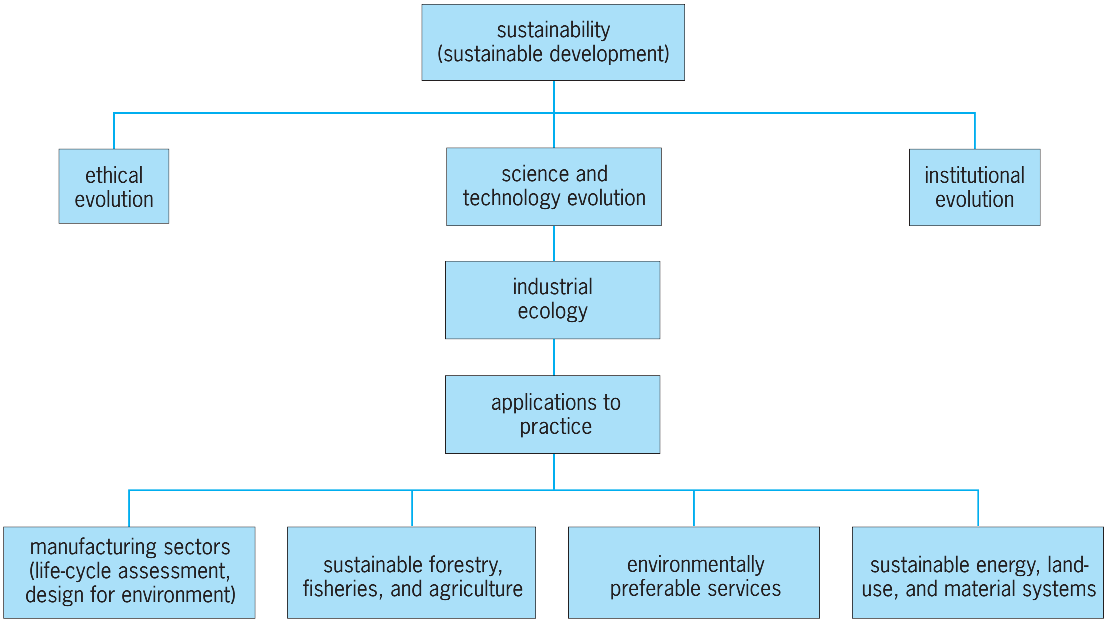


Fig. 1. Conceptual framework illustrating the relationship of industrial ecology to other concepts.

base necessary to enable progress toward sustainability. It does not replace existing disciplines and fields of study, such as engineering, environmental science, economics, or ecology; rather, it integrates them into a new and different understanding at a higher level (**Fig. 2**). In doing so, it links together results from existing, reductionist approaches, augmenting but not replacing such research. Thus, it supports a better understanding of the couplings between natural and human systems that exist in reality.

A number of different applications of, and methodologies based on, industrial ecology principles are being developed. They range from those appropriate to the firm level, such as life-cycle assessment and design for environment, to those that operate at a sector or social level, such as integrated materials management, environmentally preferable services, or sustainable fisheries. Further detail arises as each family of methodologies expands; the number of analytic tools and technologies being developed is rapidly increasing. Thus, for example, an important initiative in sustainable agriculture is integrated pest management, and intensive efforts are under way in the forestry industry to define the characteristics of sustainable forestry operations.

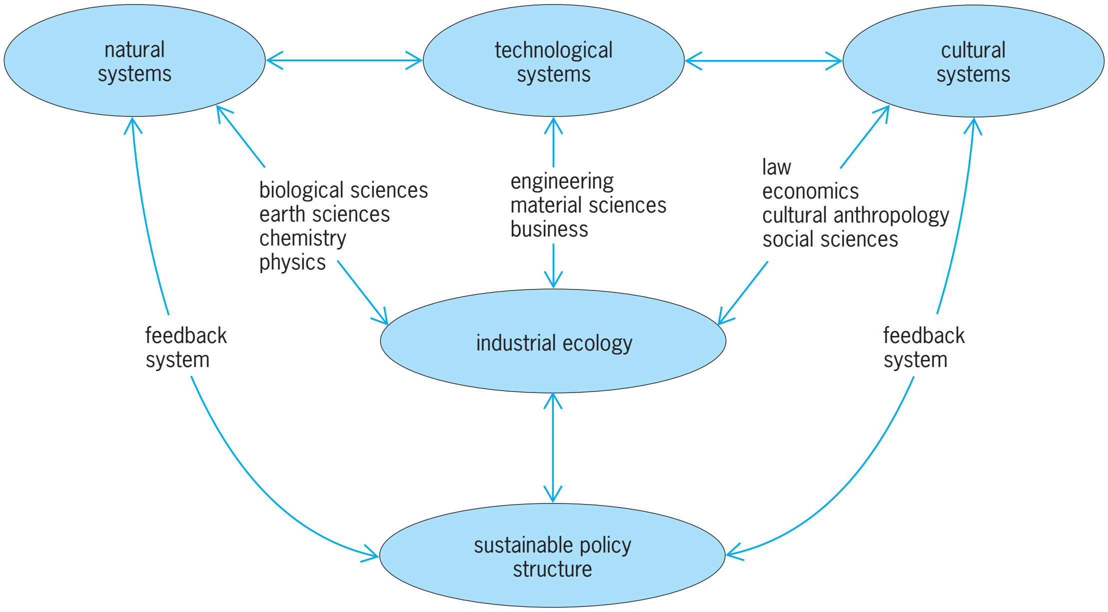


Fig. 2. Industrial ecology as an integrative multidisciplinary field. It pulls together results from existing fields of study to support an understanding of environmental perturbations in the context of complex human-natural systems.

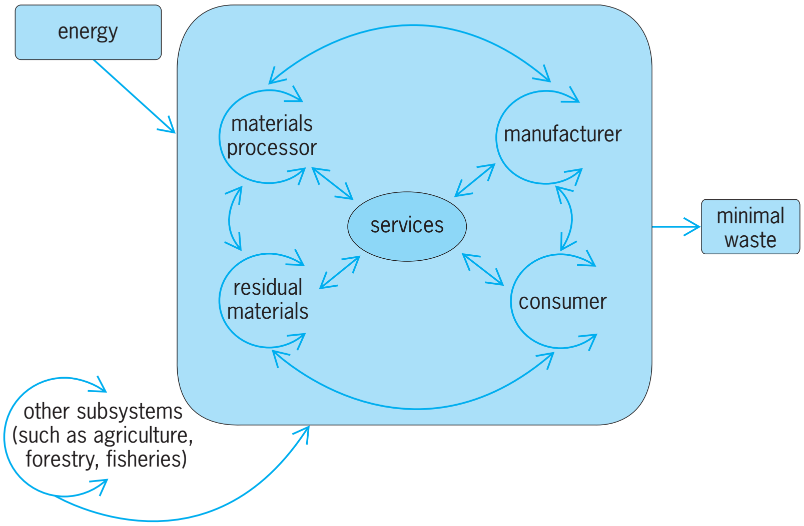


Fig. 3. Energetically open, low-waste economy. Materials are cycled through the economy at different scales and in different sectors to minimize overall waste production while quality of life is enhanced.

Principles. Some of the obvious principles of industrial ecology can be illustrated by analogy to biological communities. For example: (1) Creation of economic systems where material flows are reused rather than becoming waste is a critical element of industrial ecology (**Fig. 3**). (2) The focus is on systems and integrative analysis rather than specific elements. Key techniques which support such an approach are a focus on material stocks and flows, and energy consumption, throughout the system. (3) Industrial ecology is concerned not just with static analyses of systems, but with their responses over time, and particularly with their resilience (how well they maintain system coherence and function when stressed).

Five key concepts of industrial ecology are:

1. Design of products, processes, facilities, infrastructure, services, and technology systems so that they can be easily adapted to environmentally preferable innovation with minimal waste. Modular design of complex technologies is an example.

2. Minimization of waste production and energy consumption in all activities.

3. Use of the least toxic alternatives whenever possible, particularly when the materials may be dispersed into the environment. This principle would have argued against adding lead to gasoline, since predictably the pollutant subsequently would be dispersed throughout the environment.

4. Design of products, facilities, infrastructure, and technology systems to preserve the embedded utility of materials and energy used in initial manufacture. Thus, in many cases designs which extend the life of products and support the recycling of subassemblies or components, rather than materials, are preferable.

5. Design of physical products at all scales not just to perform their intended function but also to be used in creating other useful products at the end of their current life.

The field of industrial ecology is at present underdeveloped, and in many cases researchers are not yet even asking the right questions. The rapid development of methodologies and tools based on industrial ecology principles, the pressures for environmentally preferable behavior by firms and other institutions, and global agreements addressing complex environmental perturbations such as global climate change, however, indicate that continued growth of industrial ecology as a field of study is critical. *See* CLIMATE MODIFICATION; CONSERVATION OF RESOURCES; ECOLOGY; ECOLOGY, APPLIED; ENVIRONMENT; ENVIRONMENTAL ENGINEERING; HUMAN ECOLOGY; SYSTEMS ENGINEERING.

Braden R. Allenby

Bibliography. B. R. Allenby, *Industrial Ecology: Policy Framework and Implementation*, Prentice Hall, Upper Saddle River, NJ, 1999; R. U. Ayres and U. E. Simonis (eds.), *Industrial Metabolism*, United Nations University Press, Tokyo, 1994; T. E. Graedel and B. R. Allenby, *Industrial Ecology*, Prentice Hall, Upper Saddle River, NJ, 1995; R. Socolow et al. (eds.), *Industrial Ecology and Global Change*, Cambridge University Press, 1994.

Industrial engineering

A branch of engineering dealing with the design, development, and implementation of integrated systems of humans, machines, and information resources to provide products and services. Industrial engineering encompasses specialized knowledge and skills in the physical, social, engineering, and management sciences, such as human and cognitive sciences, computer systems and information technologies, manufacturing processes, operations research, production, and automation. The industrial engineer integrates people into the design and development of systems, thus requiring an understanding of the physical, physiological, psychological, and other characteristics that govern and affect the performance of individuals and groups in working environments.

Functions and activities. Industrial engineering is a broad field compared to other engineering disciplines. The major activities of industrial engineering stem from manufacturing industries and include work methods analysis and improvement; work measurement and the establishment of standards; machine tool analysis and design; job and workplace design; plant layout and facility design; materials handling; cost reduction; production planning and scheduling; inventory control, maintenance, and replacement; statistical quality control; scheduling; assembly-line balancing, systems, and procedures; and overall productivity improvement. Computers and information systems have necessitated additional activities and functions, including numerically controlled machine installation and programming; manufacturing systems design; computer-aided design/computer-aided manufacturing, design of experiments, quality engineering, and statistical process control; computer simulation, operations research, and management science methods; computer applications, software development, and information technology; human-factors engineering and ergonomics; systems design and integration; and robotics and automation.

The philosophy and motivation of the industrial engineering profession is to find the most efficient and effective methods, procedures, and processes for an operating system, and to seek continuous improvement. Thus, industrial engineering is utilized to help organizations grow and expand efficiently during periods of prosperity, and streamline costs and consolidate and reallocate resources during austere times. Industrial engineers, particularly those involved in manufacturing and related industries, work closely with management. Therefore, some understanding of organizational behavior, finance, management, and related business principles and practices is needed.

Functional areas. Industrial engineering can be classified into four areas: human factors and work systems, production and distribution systems, manufacturing systems, and operations research. This classification is in accordance with the various functions and activities in the design, development, and improvement of methods, operations, processes, and products.

Human factors and work systems. These combine the traditional areas of work measurement and design with the analysis and evaluation of the human element, which includes the physiological and psychological considerations in designing products and work systems. Technology continues to challenge the limitations and role of individuals and groups in production and service systems. Specifically, there is the challenge of matching capabilities such as the speed, touch, strength, and decision-making skills of humans with sophisticated computers and machine components. The influence and effects of noise, heat, humidity, lighting, and other environmental factors are dealt with in terms of the safety and welfare of the worker, as well as the evaluation of conditions for achieving optimum performance. Related issues deal with identifying and evaluating the information requirements and constraints associated with a system or design in order to achieve effective operation or service. *See* CONTROL SYSTEMS; HUMAN-FACTORS ENGINEERING; HUMAN-MACHINE SYSTEMS; INDUSTRIAL FACILITIES; INDUSTRIAL HEALTH AND SAFETY; METHODS ENGINEERING; MODEL THEORY; OPTIMIZATION; PRODUCTION METHODS; PRODUCTIVITY; SIMULATION; SYSTEMS ANALYSIS; SYSTEMS ENGINEERING; WORK MEASUREMENT; WORK STANDARDIZATION.

Production and distribution systems. These deal with the planning, scheduling, and flow of materials and information in production and manufacturing. This area includes the design, control, and implementation of production systems, including analysis procedures for determining where to locate people, machines, warehouses, and production centers; how to control and regulate the flow of parts and services to provide timely arrival of materials, acceptable quality, and customer satisfaction; scheduling and routing of operations; and maintenance and replacement procedures for ensuring cost effectiveness and overall productivity. *See* INVENTORY CONTROL; MATERIALS HANDLING; PRODUCTION PLANNING.

Manufacturing systems. These consist of organized configurations of machines, workstations, information, and processes to transform materials into products, which include parts, assemblies, and subassemblies. The processes include machining, forming, assembly, inspection and testing, and transfer. They can be achieved through various levels of automation and hierarchical control. Design and development issues involve the specification and allocation of cells, buffer storage, computer integration, degree of automation, and inspection for overall effectiveness in producing high-quality, low-cost products to meet customer expectations. *See* COMPUTER-AIDED DESIGN AND MANUFACTURING; COMPUTER-INTEGRATED MANUFACTURING; FLEXIBLE MANUFACTURING SYSTEM.

Operations research. Operations research deals with the development and application of quantitative techniques and methods for analyzing and evaluating systems that aid in making engineering and management decisions. Management science, which is more often thought to be associated with business applications, is frequently used synonymously with operations research. This area includes the development and application of analytical, experimental, and simulation models for describing complex systems and prescribing conditions for efficient and optimal performance. Mathematical programming algorithms for allocating resources; methods for analyzing queueing and facility bottlenecking problems, inventory, and machine replacement decisions; and computer computation and implementation methods are among the issues dealt with in the field of operations research. *See* ARTIFICIAL INTELLIGENCE; OPERATIONS RESEARCH; QUEUEING THEORY.

Related subdisciplines. With its extensive breadth, industrial engineering has several subfields that overlap with neighboring disciplines, such as psychology, mechanical engineering, materials science, mathematics, computer science, business, and electrical engineering. Among the larger and more recognized subdisciplines are ergonomics, human factors, manufacturing engineering, operations research, operations management, statistics, and systems engineering. In many cases the extent of overlap is so significant, particularly in relation to research programs, that industrial engineering methodologies and techniques have been adopted and assimilated by other disciplines. *See* PRODUCTION ENGINEERING.

Marlin U. Thomas

Bibliography. A. B. Badiru (ed.), *Handbook of Industrial and Systems Engineering*, 2005; H. P. Emerson and D. C. E. Naehring, *Origins of Industrial Engineering*, 1988; M. P. Groover, *Automation, Production Systems, and Computer-Aided Manufacturing*, 2d ed., 2000; G. Salvendy (ed.), *Handbook of Industrial Engineering: Technology and Operations Management*, 3d ed., 2001; K. B. Zandin and H. B. Maynard, *Maynard's Industrial Engineering Handbook*, 5th ed., 2001.

Industrial health and safety

An interdisciplinary field that focuses on preventing occupational illnesses and injuries. The disciplines of engineering, epidemiology, toxicology, medicine, psychology, and sociology provide the methods for study and prevention. In the United States, prevention efforts have resulted in a reduction of accidental work deaths per year; nevertheless, the accident toll remains high.

Occupational Hazards

Tens of thousands of occupational hazards exist. Occupational hazards can be organized by separately considering plants and equipment, the physical work environment, hazards of materials, and task demands. Significant interactions occur between these categories. For example, equipment can modify the work environment by producing noise, potentially hazardous materials, or heat, but will be hazardous only if inappropriate procedures are followed.

Plants and equipment. Plant hazards are often associated with energy sources and power transmission, processes at the point of operation, vehicles and materials-handling systems, walking and climbing surfaces, ingress-egress, and confined spaces.

Energy sources and power transmission. Potentially catastrophic results are associated with the rapid uncontrolled release of energy from high-energy sources, such as fuel supplies, pressure vessels and boilers, cylinders containing compressed gases, and electrical power supplies. Uncontrolled releases of energy may occur at the source or during power transmission to the point of operation. The nature of such hazards depends upon the energy transmitted. For example, electrical power lines pose serious dangers when they are inadvertently contacted. At high voltage levels, the electromagnetic field surrounding power lines may induce currents or cause voltage buildups. Piping systems containing fuels such as natural gas, heated liquids, steam, compressed air, or toxic materials analogously pose serious risks, if they break or are inadvertently cut by other equipment. Moving parts in mechanical power-transmission equipment are dangerous if contacted. Rotating shafts, gears, pulleys, wheels, impacts, and the turbulent flow of gases or liquids produce noise and vibration hazards. *See* ELECTRIC POWER SYSTEMS.

Point-of-operation hazards. Machinery and tools may pose a potential for abrading, cutting, shearing, or crushing injuries when people contact sharp or moving parts. Machinery may also propel flying objects and emit dusts and potentially hazardous materials, resulting in eye or other injuries. Other point-of-operation hazards include high voltages, nonionizing radiation, and hot or cold surfaces. Point-of-operation hazards usually decrease when industries shift away from manual machining tasks.

Vehicles and materials-handling systems. Collisions between vehicles or involving pedestrians are a particularly significant problem in many industries. Overhead cranes, hoists, conveyors, forklifts, carts, and other materials-handling systems can also cause impact-related injuries. Collisions and impacts cause substantial losses. *See* MATERIALS-HANDLING EQUIPMENT.

Unsafe walking or climbing surfaces. Many falls occur while people are climbing or maneuvering on elevated surfaces, such as stairs, escalators, ladders, and scaffolds, or near floor openings; but most falls involve slipping or tripping while walking, pushing, or pulling objects on level ground. The coefficient of friction between a shoe and floor surface is a primary determinant of slipping hazard. This coefficient varies between floor surfaces and shoe soles, and as a function of contaminants.

Ingress-egress and confined spaces. Inadequate or blocked passageways may impede ingress or egress from work areas or allow insufficient clearance for passing vehicles. Poorly designed ingress-egress systems can impede emergency egress from vehicles and cause falls. Confined spaces that are entered during maintenance or repair tasks, such as tanks, boilers, or ducts, can pose special ingress-egress problems, and may concentrate hazardous vapors, fumes, and gases.

Physical work environment. Hazards in the physical work environment include vibration and noise, thermal extremes, pressure extremes, and ionizing or nonionizing radiation.

Vibration and noise. Excessive vibration and noise is a potential safety problem in vehicles and many work environments. Whole-body vibration at certain resonant frequencies is annoying and fatiguing to workers and drivers and may cause chronic pain. Localized vibration produced by some handtools can interfere with blood circulation, resulting in Raynaud's phenomenon, damage to the soft tissues of the hand, bursitis, or arthritis.

Noise becomes a safety problem at certain levels and frequencies. The most commonly used scale [the adjusted decibel (dBA) scale] attenuates low frequencies not normally used for speech communication. Even at low levels, noise can annoy or distract workers. At moderate levels of 80–90 dBA, noise interferes with communication and causes hearing loss in susceptible people. Higher levels (above 85 dBA) significantly increase the chance of hearing loss. The effects of long-term exposure to noise are cumulative and nonreversible. In the United States the Occupational Safety and Health Act specifies allowable levels and durations of exposure to noise (**Table 1**). The act also requires that employers monitor noise levels, perform audiometric tests, furnish hearing protection, and maintain exposure records whenever employee noise exposure equals or exceeds an 8-h time-weighted average level of 85 dBA. *See* ACOUSTIC NOISE; AUDIOMETRY; HEARING (HUMAN); HEARING IMPAIRMENT.

Thermal extremes. Excessively hot or cold working environments can adversely affect workers. Working at high temperatures can result in heat cramps, heat exhaustion, or heat stroke. Such effects become more likely when the work is strenuous, the exposure is prolonged, the humidity is high, or the

TABLE 1. Permissible noise exposure

Duration per day	Slow response, decibel level, adjusted decibels (dBA)
8	90
6	92
4	95
3	97
2	100
1$\frac{1}{2}$	102
1	105
$\frac{1}{2}$	110
$\frac{1}{4}$ or less	115

workers are using respirators or wearing protective clothing that interferes with normal evaporative cooling. Working at low temperatures can result in localized tissue damage, due to restriction of circulation, or hypothermia at temperatures well above freezing; Frostbite can occur below freezing. Air movement increases the potential for frostbite and hypothermia. Moisture also increases the rate of cooling and the risk associated with cold temperatures. *See* COMFORT TEMPERATURES; HYPOTHERMIA.

Pressures. Environments above or below normal atmospheric pressure can be a significant hazard. Hyperbaric pressures are found in diving, underwater, or in tunneling operations where compressed gas is used to prevent the entry of water. Decompression injuries may occur in such tasks, when gases expand in body cavities or when nitrogen dissolved in the bloodstream or other tissues changes to gaseous form. One solution is to slowly carry out the decompression process. A second is to provide mixtures of compressed oxygen and helium, rather than oxygen and nitrogen (helium is less soluble in blood and other tissues). Hypobaric pressures are found at high altitudes. Decompression rapid enough to cause pain or injury to the inner ear and sinuses may occur when airplanes rapidly lose cabin pressure or reduce altitude by a few thousand feet. *See* AIR PRESSURE; DECOMPRESSION ILLNESS; DIVING.

Ionizing and nonionizing radiation. Radioactive materials that emit alpha and beta particles, neutrons, x-rays, or gamma rays are sources of ionizing radiation. X-rays and radioactive materials are commonly used in medicine, chemical laboratories, and other industries to scan or obtain information from within living or inert objects. They are also used to generate power. Ionizing radiation has sufficient energy to ionize components of human cells. Radioactive materials consequently pose a potential hazard if excessive amounts are ingested, inhaled, or otherwise introduced into the body. Sources of alpha particles pose a hazard only when within the body. Sources of beta particles are capable of burning the skin but are primarily a hazard when within the body. Sources of neutrons, x-rays, and gamma rays pose hazards both within and outside the body. Control measures include limits on exposure time, adequate separation distance from the source, and use of barriers. *See* RADIATION BIOLOGY; RADIATION INJURY (BIOLOGY).

Forms of nonionizing radiation include visible, ultraviolet, infrared, and microwave radiation. Visible light is focused by the lens of the eye onto the retina, while the other forms are beyond the visible spectrum. High-intensity visible light may burn the retina, resulting in color blindness or blind spots. Ultraviolet light can seriously burn human skin and the cornea of the eye. Repeated exposure has been implicated as a cause of skin cancer and cataracts. Welding equipment, lasers, and the Sun emit potentially dangerous intensities of visible and ultraviolet light. Infrared radiation is emitted by hot objects and may excessively heat workers, resulting in heat exhaustion, heat stroke, or heat cramps. At high levels, it can burn the skin and eyes and cause cataracts. Microwave radiation is emitted by equipment such as radar, microwave ovens, and communication apparatus. Microwaves can cause serious tissue damage by excessive heating of exposed body parts and inductive heating of metals. *See* ELECTROMAGNETIC RADIATION; LASER; MICROWAVE; RADAR.

Hazards of materials. Materials used in industrial processes vary greatly in nature and form. Mists, vapors, gases, liquids, dusts, and fumes from certain materials may be hazardous. Some materials pose fire and explosion hazards. Others are chemically or biologically active when they contact or enter the human body. Even chemically inert materials can cause injuries or illness.

Fire and explosion hazards. Fires and explosions are major hazards of certain materials. Numerous materials can burn and create fire hazards. Airborne gases, vapors, fumes, or dusts are particularly dangerous, since they can travel significant distances to reach ignition sources and often burn explosively. Beyond the hazards of heat and potentially explosive forces, combustion of materials can result in highly toxic products. Fires also deplete the oxygen content of air, creating the potential for suffocation. *See* COMBUSTION; EXPLOSIVE; FIRE TECHNOLOGY.

Chemical and biological hazards. Certain chemicals cause immediate or delayed injury or even destruction of various organs upon bodily contact or entry. Biologically active agents, such as bacteria, viruses, fungi, mycoplasma, or parasitic organisms, cause disease. Chemicals and biologically active agents can enter the body through the skin, by ingestion, or by inhalation. Some agents readily pass through unbroken skin; others may enter through skin abrasions or cuts. Workers may ingest contaminated food or fluids while eating or smoking without first washing their hands. Dusts, fumes, mists, vapors, gases, and biologically active agents may also be inhaled.

The effects of chemicals are related to toxicity, particle size, and mode of entry into the body. Several categories of toxic chemical effects have been determined. Asphyxiation refers to the effects of chemicals that cause an inadequate supply of oxygen to be delivered through the blood to organs within the body. Simple asphyxiants dilute the air that is breathed so that not enough oxygen is present to support life. Chemical asphyxiants interfere with the ability of blood to deliver oxygen. *See* HYPOXIA.

Irritation occurs when contact with a chemical causes an inflammatory or allergic response by the body. Fine particulate matter or dusts can cause irritation of the respiratory tract and lungs, even when the matter is nontoxic or chemically inert. Symptoms of irritation include an increased flow of blood to the location of contact, swelling, and the activation of immune system response. Severe cases may involve dermatitis, extensive blistering, pulmonary edema, or interference with breathing. *See* ALLERGY; INFLAMMATION; RESPIRATORY SYSTEM DISORDERS.

Corrosive or chemical burns occur when a chemical agent reacts with the contacted surface to break down the tissues. Systemic poisons interfere with the metabolic functions of tissues and often affect only particular organs; neurotics, depressants, anesthetics, stimulants, and hypnotics target the nervous system; and carcinogens and mutagens cause various forms of cancer. *See* MUTAGENS AND CARCINOGENS; TOXICOLOGY.

Task demands. The task performed by a worker can be hazardous. Lifting, pushing, pulling, and other physical activity can cause injury when applied or reactive forces, pressures, or torques exceed the tolerance of the body. Repeated performance of manual tasks over prolonged periods, excessive reaches, twisting motions, rapid movements, and postures that concentrate forces can significantly increase the risk of injury. Tasks that are stressful or monotonous can also contribute to human error. Changes in work conditions requiring deviations from ordinary routines, such as when equipment is being repaired, are particularly likely to increase the chance of errors.

Safety Activities

Safety and health activities performed within industry include identifying and analyzing hazards, ensuring compliance with standards and codes, implementing controls, and monitoring controls.

Hazard identification and analysis. A fundamental safety and health activity is to identify potential hazards and then to analyze them in terms of severity and probability. This process allows the cost of control measures to be compared with expected loss reduction and helps justify choices between control alternatives.

Hazard identification is guided by past experience, codes and regulations, checklists, and other sources. This process can be organized by separately considering each step in the making of a product (**Fig. 1**). Numerous complementary hazard analysis methods are available, including failure modes and effects analysis, work safety analysis, human error analysis, and fault tree analysis. Failure mode and effects analysis systematically documents the effects of malfunctions on work sheets that list the components of a system, their potential failure modes, the likelihood and effects of each failure, and potential countermeasures. Work safety analysis and human error analysis are related techniques that organize the analysis around tasks rather than system components. This process involves an initial division of tasks into subtasks. For each subtask, potential effects of product malfunctions and human errors are then documented. Fault tree analysis takes an approach that begins with a potential accident and then works down to its fundamental causes. Fundamental causes may be system malfunctions, human errors, or ordinary nonmalfunction states. Probabilities are often assigned to the fundamental causes, allowing the probability of accidents to be calculated. *See* OPERATIONS RESEARCH; RISK ASSESSMENT AND MANAGEMENT; SYSTEMS ANALYSIS.

Compliance with standards and codes. Determining which standards, codes, and regulations are

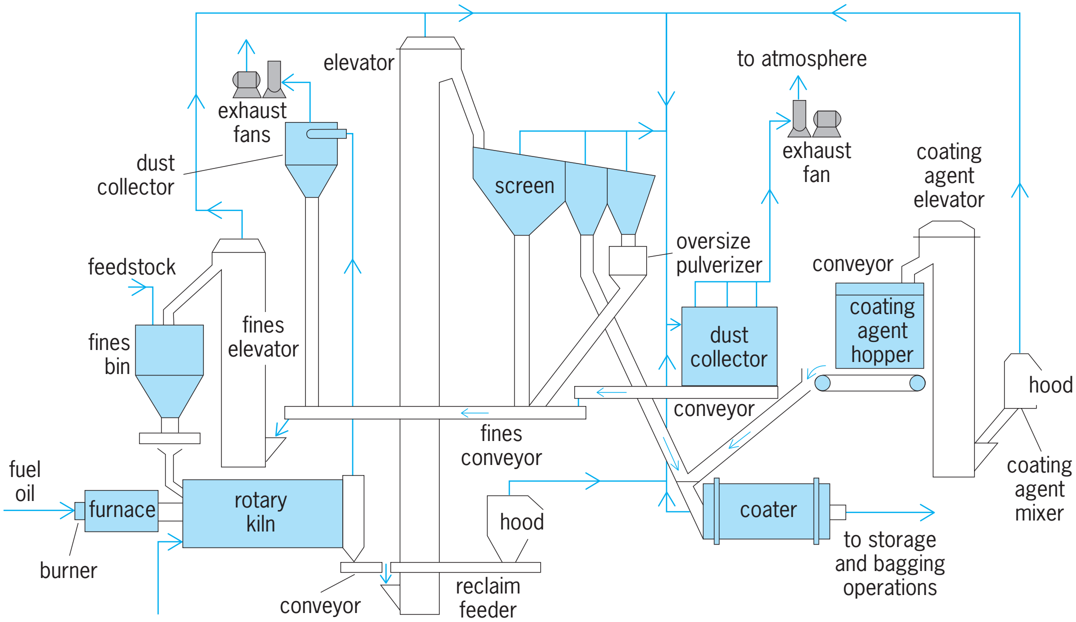
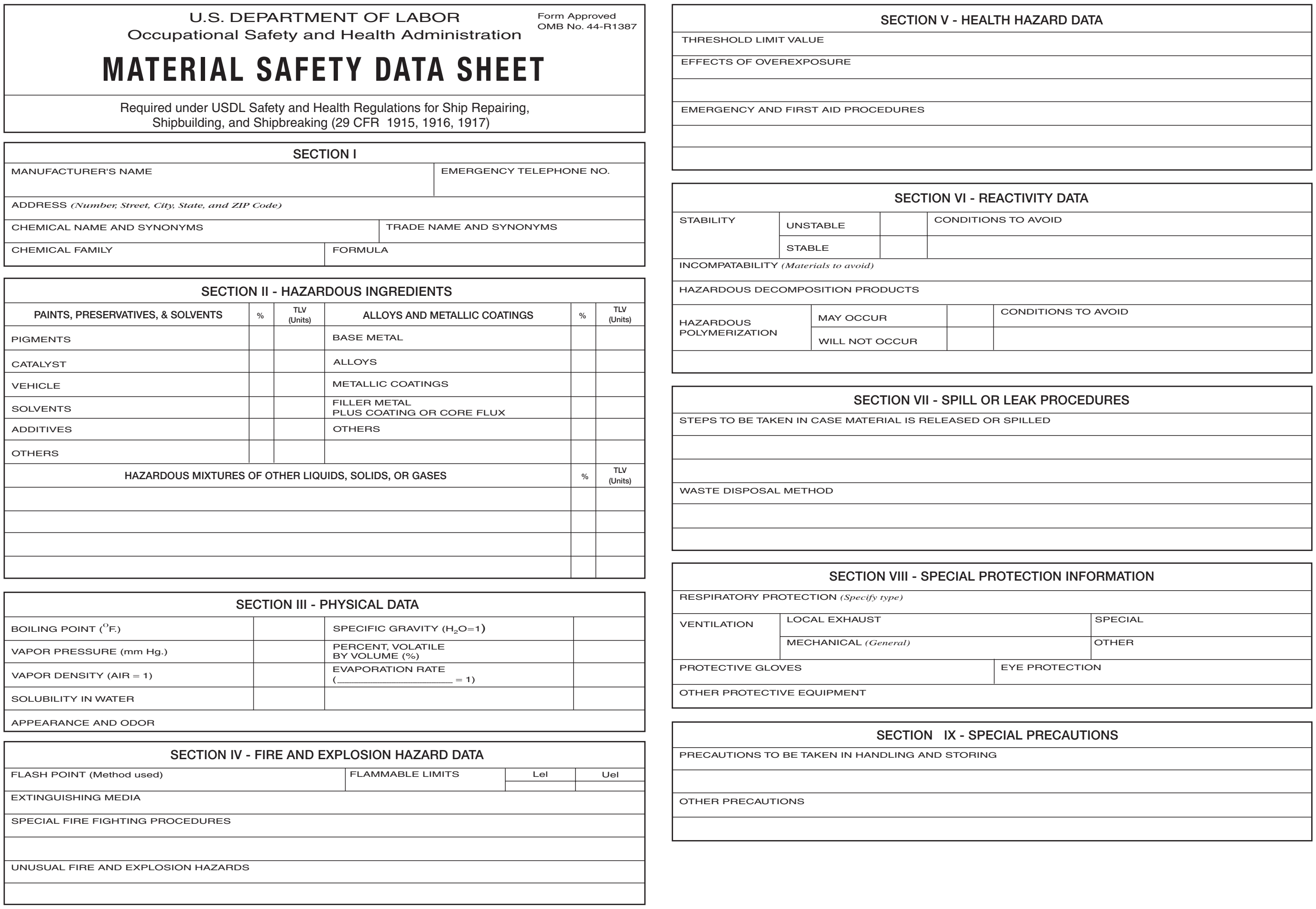


Fig. 1. Process flow sheet showing the stepwise introduction of raw materials and the product of each step. The extent of chemical or physical hazards that may arise out of any step in the operation should be determined.



U.S. DEPARTMENT OF LABOR
Occupational Safety and Health Administration
Form Approved
OMB No. 44-R1387

MATERIAL SAFETY DATA SHEET

Required under USDL Safety and Health Regulations for Ship Repairing,
Shipbuilding, and Shipbreaking (29 CFR 1915, 1916, 1917)

SECTION I

MANUFACTURER'S NAME
EMERGENCY TELEPHONE NO.
ADDRESS *(Number, Street, City, State, and ZIP Code)*
CHEMICAL NAME AND SYNONYMS
TRADE NAME AND SYNONYMS
CHEMICAL FAMILY
FORMULA

SECTION II - HAZARDOUS INGREDIENTS

PAINTS, PRESERVATIVES, & SOLVENTS	%	TLV (Units)	ALLOYS AND METALLIC COATINGS	%	TLV (Units)
PIGMENTS			BASE METAL		
CATALYST			ALLOYS		
VEHICLE			METALLIC COATINGS		
SOLVENTS			FILLER METAL PLUS COATING OR CORE FLUX		
ADDITIVES			OTHERS		
OTHERS					

HAZARDOUS MIXTURES OF OTHER LIQUIDS, SOLIDS, OR GASES	%	TLV (Units)

SECTION III - PHYSICAL DATA

BOILING POINT (^{O}F.)		SPECIFIC GRAVITY (H_2O=1)	
VAPOR PRESSURE (mm Hg.)		PERCENT, VOLATILE BY VOLUME (%)	
VAPOR DENSITY (AIR = 1)		EVAPORATION RATE (____________ = 1)	
SOLUBILITY IN WATER			
APPEARANCE AND ODOR			

SECTION IV - FIRE AND EXPLOSION HAZARD DATA

FLASH POINT (Method used)
FLAMMABLE LIMITS Lel Uel
EXTINGUISHING MEDIA
SPECIAL FIRE FIGHTING PROCEDURES
UNUSUAL FIRE AND EXPLOSION HAZARDS

SECTION V - HEALTH HAZARD DATA

THRESHOLD LIMIT VALUE
EFFECTS OF OVEREXPOSURE
EMERGENCY AND FIRST AID PROCEDURES

SECTION VI - REACTIVITY DATA

STABILITY	UNSTABLE		CONDITIONS TO AVOID
	STABLE		

INCOMPATABILITY *(Materials to avoid)*
HAZARDOUS DECOMPOSITION PRODUCTS

HAZARDOUS POLYMERIZATION	MAY OCCUR		CONDITIONS TO AVOID
	WILL NOT OCCUR		

SECTION VII - SPILL OR LEAK PROCEDURES

STEPS TO BE TAKEN IN CASE MATERIAL IS RELEASED OR SPILLED
WASTE DISPOSAL METHOD

SECTION VIII - SPECIAL PROTECTION INFORMATION

RESPIRATORY PROTECTION *(Specify type)*

VENTILATION	LOCAL EXHAUST	SPECIAL
	MECHANICAL *(General)*	OTHER

PROTECTIVE GLOVES
EYE PROTECTION
OTHER PROTECTIVE EQUIPMENT

SECTION IX - SPECIAL PRECAUTIONS

PRECAUTIONS TO BE TAKEN IN HANDLING AND STORING
OTHER PRECAUTIONS

Fig. 2. Sample Material Safety Data Sheet, used to obtain information from the manufacturer. (*Occupational Safety and Health Administration*)

relevant and then ensuring compliance are essential health and safety activities. Consensual standards developed by the American National Standards Institute (ANSI) are particularly important; they address topics such as chemical labeling, material safety data sheets, personal protective equipment, workplace signs and product labels, and pressure vessels. Other widely applicable standards developed in the United States by nongovernmental groups include the National Electric Code, the Life Safety Code, and the American Council of Industrial and Governmental Hygienists guidelines for threshold limit values. Thousands of consensual standards contain safety provisions.

In the United States, the best-known governmental standards are the general industry standards specified by the Occupational Safety and Health Administration (OSHA). OSHA also specifies standards for the construction, maritime, and agriculture industries. Other standards include those specified by the Environmental Protection Agency (EPA) on disposal and cleanup of hazardous materials, the Federal Aviation Administration (FAA) standards on worker safety in air travel, the Federal Highway Administration (FHWA) standards regarding commercial motor carriers, the Mine Safety and Health Administration (MSHA) standards for mine workers, the Nuclear Regulatory Commission's and Department of Energy's standards regarding employees working with radioactive materials, and the U.S. Coast Guard standards regarding safety of workers on tank and passenger vessels. State and local governments may also implement safety and health standards.

The scope of governmental regulations is illustrated by the OSHA general industry standard. Besides specifying a large set of detailed and mandatory health and safety design standards, OSHA addresses noncovered hazards with a general duty clause that places an obligation on employers to provide a workplace free from recognized hazards. OSHA also requires that industries implement comprehensive hazard communication programs for workers involving labeling, material safety data sheets, and employee training (**Fig. 2**). OSHA enforces these standards by conducting inspections and imposing fines if they are not met. Employers are required to maintain records of work-related injuries and illnesses and to prominently post both an annual summary of injuries and illness and notices of noncompliance with standards. Other functions of OSHA include monitoring the performance of state agencies which implement health and safety standards and coordinating such activity with that of other regulatory agencies.

Control methods. Methods of controlling or eliminating hazards include plant or process design; job design; employee selection, training, and supervision; personal protective equipment; and warnings.

Plant or process design. Plant and process design can eliminate or control hazards in numerous ways. Eliminating hazards at their source by substituting less hazardous materials, equipment, or processes is an ideal solution, if feasible. For example, hazardous mists can be eliminated by brush painting instead of spray painting. Processes can be modified to eliminate hazards at the source, such as wetting-down grinding operations to minimize dust production. Another solution is to eliminate hazardous jobs through automation. For example, robot welders and painters have eliminated hazards in the automobile industry. *See* AUTOMATION; DUST AND MIST COLLECTION; HAZARDOUS WASTE; ROBOTICS.

Eliminating hazards, however, may be technically infeasible. An alternative approach is to isolate them, thereby reducing the number of potentially exposed workers; the conflicts between processes, such as welding and paint spraying operations; the potential contact with power transmission equipment, such as power lines; and the catastrophic potential of high-energy systems. Isolation may be attained through barriers and enclosures that block or absorb hazardous flows of energy or material. Plant layouts can contribute to this goal by spatially separating hazards and susceptible structures. Barriers such as fire doors and fire walls, along with fire suppression and warning systems, provide time to escape and respond to fires. Process enclosures prevent the release of hazardous chemicals, noise, heat, or other hazards into the workplace. Appropriate passageways separate workers from vehicles and hazardous processes. Under emergency conditions, they also speed egress of workers and access by emergency equipment and personnel. Detailed provisions regarding passageways and emergency exits include appropriate passage dimensions, markings, signage, lighting, stairs, walking surfaces, door locations, and door designs. *See* AIR FILTER; INDUSTRIAL ENGINEERING; INDUSTRIAL FACILITIES.

The removal or dilution of airborne contaminants through local or dilution ventilation is another important control measure. Local ventilation removes contaminants near their source, often from within enclosed processes, making it more effective than dilution ventilation. OSHA specifies extensive ventilation requirements for tasks producing airborne contaminants. *See* VENTILATION.

Job design. Job design involves the specification of appropriate methods, tools, and equipment. Focus may also be on safety procedures and safety devices. Each aspect of job design can play a critical role in eliminating hazards, preventing contact with hazards, and minimizing effects of contact.

In some cases, hazards can be entirely eliminated through job design. For example, the need for heavy lifting, pushing, or pulling may be eliminated by providing hoists, conveyors, or other materials-handling equipment. Automation may eliminate the need for repetitive manual operations in assembly tasks. More often, job design reduces hazards. For example, providing grounded tools, nonconductive mats, and nonconductive ladders reduces, but does not eliminate, electrocution hazards. Job design can also keep physical demands within reasonable bounds through job rotation, rest intervals, and ergonomic guidelines. Among such guidelines are those that describe limits on lifted weights as a function of load location, lifting height, task duration, and lifting frequency. Guidelines oriented toward preventing carpal tunnel syndrome specify acceptable wrist orientations and exerted forces when using keyboards, hand tools, and other devices. Guidelines for preventing falls in climbing systems specify appropriate handholds, steps, or rungs. Guidelines for preventing falls in walking, lifting, pushing, or pulling tasks specify slip-resistant work surfaces, appropriate footwear, and limits on exerted forces.

Contact with hazards often can be prevented by modifying tools and equipment. One approach is to modify contact surfaces by eliminating sharp edges, points, rough surface finishes, pinch points, and temperature extremes. Guards also can prevent workers from contacting hazardous components of equipment or tools, but are less effective than eliminating the hazard since they can be removed by workers.

Other approaches address unsafe behavior due to perceptual errors and memory lapses. Examples include delineating risk zones with barriers or markings, providing unobstructed vision, and reducing levels of noise. Inadvertent activation of equipment and entry into risk zones can be minimized by providing temporary barriers, interlocks, and lockouts. Errors can also be reduced by providing appropriate displays and controls, reducing demands on short-term memory, and lowering the cognitive complexity of tasks. Job performance aids, such as checklists, instruction manuals, and computer programs, can be particularly useful for preventing errors when performing nonroutine tasks, such as maintenance, repair, and troubleshooting. *See* HUMAN-MACHINE SYSTEMS.

Employee selection, training, and supervision. Modification of the employee through employee selection, training, and supervision is necessary in almost all occupations. Selecting employees with the necessary physical and cognitive abilities is essential. For example, airplane pilots must have good visual acuity, the safe use of hand tools requires some minimal level of dexterity, and lifting tasks require some minimal level of strength.

Training workers and supervisors is important. A training program must teach workers what the hazards are and their severity, how to identify and avoid them, and what to do after exposure. Trained workers should follow safe procedures, including steps to take after exposure, such as washing and rinsing after skin or eye contact with hazardous materials; methods of proper disposal of contaminated clothing, hazardous chemicals, or biologically active agents; necessary maintenance activities; and housekeeping procedures, such as cleaning up spills or accumulations of materials in the work environment. Housekeeping procedures are particularly essential to prevent falls, fires, and exposure to hazardous materials.

Supervision, as the final element related to modification of the employee, should focus on consistently reinforcing safe procedures.

Personal protective equipment. Personal protective equipment, such as gloves, respirators or masks, protective clothing, hearing protectors, eye protectors, foot protectors, helmets, seat belts, and safety harnesses, may be useful when other control methods are inadequate or infeasible. Training in proper use of personal protective equipment is important and supervision may be necessary to ensure that the equipment is used properly when needed. Plant and task layouts should provide convenient access to personal protective equipment. *See* RESPIRATOR.

Warnings. Warnings are the method of last resort, and they should be viewed only as a supplement to proper training and supervision. Warnings are of primary value as a means of alerting knowledgeable users to follow an established safe behavior pattern. When established behavior patterns are unsafe, supervision becomes the most important option.

Monitoring controls. Accident investigations, plant inspections, and monitoring are complementary ways of ensuring that implemented control strategies are fulfilling their intended function. They can uncover deficiencies in existing controls and help formulate needed changes.

Accident investigation. Accident investigation provides a means of determining why existing control methods failed. Accident investigations vary from simple documenting of what happened on a supervisor's report to in-depth studies involving governmental agencies and experts in forensics, medicine, human-factors engineering, and other fields. Supervisors' reports document who was injured, the nature and extent of injury, location and time of the accident, and what was being done at the time of the accident. The reports also include a narrative description of cause and corrective actions. Accident investigations almost always reveal multiple causes for accidents. An accident investigation should document violations of safety standards and operating procedures, equipment malfunctions, errors, and unsafe conditions. More sophisticated investigations focus on the root causes of equipment malfunctions and human errors, such as management oversights (inadequate facilities, training, supervision, emergency procedures and so on), complex interactions between processes, or error-provocative situations.

Plant inspections. Plant inspections are usually performed on a periodic basis to determine whether equipment, machinery, and tools are within design specifications. Routine quick walkthroughs by inhouse safety engineers and supervisors can uncover unsafe conditions and violations of operating procedures. Inspections should focus on environmental control systems, such as ventilation systems; pollution control systems; walking and climbing surfaces; tools, machinery, and materials-handling systems; accumulations of flammable or toxic materials; violations of operating procedures, such as failure to use personal protective equipment; disabling or removal of guards; unsafe operations of equipment and vehicles; lack of maintenance; failures to lock out equipment during repair or maintenance; and smoking in restricted areas.

Monitoring. Monitoring is a continuing program of observation, measurement, and judgment of the exposure of workers to chemical and physical hazards of the work environment (**Table 2**). Monitoring is essential to determine whether control methods are successfully reducing exposure. The four most

TABLE 2. Sampling procedures for industrial hygiene surveys

Chemical hazards (airborne)		Physical hazards	
Gases and vapors	May be determined by use of approved calibrated field indicator tubes yielding direct readings	Pressure	May be measured barometrically
	May be collected in containers or absorbed on charcoal for laboratory evaluation	Temperature	May be measured by thermometer, thermocouple, or radiometer; determination of heatstress, however, requires in some form the measurement of evaporation rate; usually heat stress is inferred from humidity and air velocity
	Organic vapor may be absorbed on charcoal or determined chromatographically	Ionizing radiation	May be measured by survey meter, personal dosimetry, or film badge techniques
Fumes and mists	May be absorbed and measured in the field		
	May be absorbed and evaluated in the laboratory	Noise levels	May be measured with sound-level meters or octave-band analyzers; vibration may be determined with additional sound-level equipment
	May be collected on filter media and analyzed in the laboratory		
Dusts	May be collected by a personal air sampler and fractionated into respirable size by a cyclone separator, and the fractions weighed to determine the concentration	Nonionizing radiation	May be measured by a number of direct-reading meters
	May be collected on an open-faced filter and weighed		
	May be collected in an appropriate manner and counted		

common methods are environmental, personal, biological, and medical monitoring. Environmental monitoring is performed at locations where employees are likely to be exposed to hazards. Noise levels, radiation levels, concentrations of airborne contaminants, and temperatures are routinely measured to determine risks. Technically advanced systems may use remote sensors to detect transient and potentially hazardous changes in the work environment. Certain systems may respond to unacceptable deviations by giving warnings, triggering fire suppression systems, or adjusting flow rates in ventilation systems.

Personal monitoring entails placing a dosimeter on a worker as close as possible to the port of entry. Thus, the amount of exposure that the employee receives over the course of a work interval can be measured. For example, a dosimeter placed in the breathing zone can be used to estimate the amount of hazardous material inhaled. Biological monitoring involves the measurement of hazardous materials or metabolic by-products of hazardous materials in bodily fluids, tissues, hair, or expired air. Examples include the measurement of lead, fluoride, cadmium, or mercury in blood or urine, or of phenol in urine to measure excessive benzene absorption. Medical monitoring is a related technique used to measure health effects on workers after exposure; an example is the routine chest x-rays of workers. *See* CLINICAL PATHOLOGY; DOSIMETER; HUMAN-FACTORS ENGINEERING.

Mark R. Lehto

Bibliography. D. B. Chaffin et al., *Occupational Biomechanics*, 4th ed., 2006; L. J. DiBerardinis (ed.), *Handbook of Occupational Safety and Health*, 2d ed., 1998; H. W. Heinrich, D. Peterson, and N. Roos, *Industrial Accident Prevention*, 5th ed., 1980; M. R. Lehto and J. M. Miller, *Warnings*, vol. 1: *Fundamentals, Design, and Evaluation Methodologies*, 1986; National Institute for Occupational Safety and Health, *Work Practices Guide to Manual Lifting*, NIOSH 81–122, 1981; National Safety Council, *Accident Prevention Manual for Industrial Operations*, vol. I: *Administration and Programs*, vol. II: *Engineering and Technology*, 9th ed., 1988; G. Salvendy (ed.), *Handbook of Industrial Engineering*, 3d ed., 2002; G. A. Taylor, K. Easter, and R. Hegney, *Enhancing Occupational Safety and Health*, 2004.

Industrial meteorology

The application of meteorological information to industrial, business, or commercial problems. Generally, industrial meteorology is a branch of applied meteorology, which is the broad field where weather data, analyses, and forecasts are put to practical use. In recent years, the term "private sector meteorology" has taken on the broader context of traditional industrial meteorology, expanding to include the provision of weather instrumentation/remote sensing devices, systems development and integration, and various consulting services to government and academia as well as value-added products and services to markets in industry (such as media, aviation, and utilities). Some areas in which industrial meteorology may be applied include environmental health and air-pollution control, weather modification, agricultural and forest management, and surface and air transportation. *See* AERONAUTICAL METEOROLOGY; AGRICULTURAL METEOROLOGY; METEOROLOGICAL INSTRUMENTATION.

Specific examples of the uses of industrial meteorology include many in the public sphere. For example, electric utilities need hourly predictions of temperature, humidity, and wind to estimate system load. In addition, they need to know when and where thunderstorms will impact their service area, so that crews can be deployed to minimize or correct disruptions to their transmission and distribution systems. Highway departments need to know when and where frozen precipitation will affect their service areas so that crews can be alerted, trucks loaded with sand and salt, and, if necessary, contractors hired to assist. Since a few degrees' change in temperature, or a slight change in intensity of snow or ice, determines the type of treatment required, early prediction and close monitoring of these parameters are critical.

Agricultural enterprises, from farmers to cooperatives to food manufacturers, rely on precise weather information and forecasts. Weather is the single most important factor in determining crop growth and production. Thus, monitoring and prediction of drought, floods, heat waves, and freezes is of extreme importance.

Professionals involved with the meteorological aspects of air pollution are generally concerned with the atmospheric transport, distribution, transformation, and removal mechanisms of air pollutants. They are often called upon to evaluate the effectiveness of pollution control technologies or regulatory (policy) actions used to achieve and maintain air-quality goals.

History. The genesis of industrial meteorology began at the end of World War II. During the war, meteorologists within the armed forces were responsible for providing site- and time-specific forecasts to support military operations. With the introduction of radar and other observing and forecasting methods, meteorologists became problem solvers for their military clients. After the war, some of these meteorologists remained in the field and developed private sector businesses. Through the 1950s and 1960s the primary service provided by these companies was weather forecasting: short- and medium-term predictions of site-specific weather conditions.

The 1970s and 1980s saw a substantial increase in the availability of real-time weather information, fueled by increases in computing power and communications technology, such as radar and satellite imagery, numerical model output, and surface observations. Eventually the traditional methods of transmitting these data, such as analog phone lines and slow-speed teletypes, became insufficient. High-speed computer-to-computer connections, satellite distribution, and local (computer) processing replaced the hand-plotted paper weather maps, wet-paper facsimile machines, and miles of teletype paper rolls that had been in use since World War II. This achievement belonged to a new segment of industrial meteorology—the value-added weather information industry. Seeing the need to capture the megabytes of weather data available every minute and transform them into usable information, the value-added industry blossomed from a few small start-up companies, in the late 1970s, to many organizations with hundreds of employees and annual sales in the tens of millions of dollars by the mid-1990s.

Value-added meteorology. These companies absorb vast amounts of digital data provided by government and other sources, such as the United Nations' World Meteorological Organization and aircraft pilot reports, and turn it into useful information for their clients. Value-added products include animated graphics, interactive model manipulation, high-powered workstations with associated software, and Internet site content. The value-added sector also provides satellite transmission and delivery of its products, customer service and training, and other user-specific services. Similarly, systems integration organizations design, develop, install, and maintain customized, high-powered meteorological processing and display systems for clients in government, industry, research, and academia. Today, the value-added weather information and systems integrations sectors are substantially larger (in personnel employed and sales volume) than the traditional weather forecasting and consulting companies, though these latter sectors remain strong.

Consulting and media services. At present, demand for personalized consulting services continues to be strong. The American Meteorological Society (AMS) has certified over 600 consulting meteorologists since the inception of the Certified Consulting Meteorologist (CCM) program in 1957. Certified Consulting Meteorologists provide services in a number of areas, including the provision of expert testimony in legal cases (forensic meteorology). The National Council of Industrial Meteorologists was founded in 1968 to further the development and expansion of the professional practice of industrial meteorology.

Media services are also an important part of industrial meteorology. It is estimated that about 75% of the weathercasters on television and radio in the United States are professional meteorologists. Since 1959 the American Meteorological Society has awarded its Seal of Approval to over 150 individuals in radio and over 1000 in television. The public's appetite for frequent, reliable weather information is evidenced by the popularity of full-color newspaper weather pages and a cable television national weather network, both products of the private sector.

Weather forecasting. With the significant improvements in short- and long-range forecasting capabilities in the 1990s, the insurance and financial communities have created risk-hedging "weather derivative contracts" for weather-sensitive industries such as utilities and agriculture. Private-sector meteorologists are assisting in the development of these markets through the refinement and delivery of temperature data and seasonal forecasts. *See* WEATHER FORECASTING AND PREDICTION.

In the United States, the federal government provided weather forecasting and prediction services for decades through local offices of the National Weather Service (NWS). Surplus demand for these services, combined with the desire of many in the private sector to provide the services themselves, led to the formulation of a mixed economic approach to weather forecasting. In 1989 the NWS (formerly the U.S. Weather Bureau, and now an agency of the National Oceanic and Atmospheric Administration within the U.S. Department of Commerce) developed Circular A-06, *Policy and Guidelines Governing National Weather Service and Private Sector Roles*.

The primary mission of the NWS, according to these policy guidelines, is the protection of life and property and the enhancement of the national economy. Hence, the basic functions of the NWS are the provision of forecasts and warnings of severe weather, flooding, hurricanes, and tsunami events; the collection, exchange, and distribution of meteorologic, hydrologic, climatic, and oceanographic data and information; and the preparation of hydrometeorological guidance and core forecast information. The NWS is the single official voice when issuing warnings of life-threatening situations and is the source of a common national hydrometeorological information base. This national information base forms an infrastructure on which the private sector can build and grow. *See* HYDROMETEOROLOGY.

The NWS does not compete with the private sector when a service is currently provided or can be provided by commercial enterprises, unless otherwise directed by applicable law. The private weather industry is ideally suited to put the basic data and common hydrometeorological information base from the NWS into a form, and with the necessary detail, that is most useful to specific weather and water resources–sensitive users. Todd S. Glickman

Bibliography. J. R. Fleming (ed.), *Historical Essays on Meteorology 1919–1995*, pp. 417–442, American Meteorological Society, Boston, 1996; T. S. Glickman, *Glossary of Meteorology*, 2d ed., American Meteorological Society, Boston, 2000; D. D. Houghton, *Handbook of Applied Meteorology*, John Wiley, New York, 1985; D. D. Houghton, Results of the 1995 AMS Private Sector Survey, *Bull. Amer. Meteorol. Soc.*, 77(2):325–333, 1996.

Industrial microbiology

A field concerned with the development of technologies to control and manipulate the growth and activities of selected biological agents to create desirable products and economic gain or to prevent economic loss. In addition to bacteria and yeasts, animal and plant cell cultures are now used to produce sophisticated products such as monoclonal antibodies, immunomodulating compounds, and complex plant metabolites.

The fermentation of foods and beverages was the earliest application of industrial microbiology. Although such products have been consumed for thousands of years, only in the nineteenth century was microbial activity associated with the fermentation process. Soon after that discovery, microorganisms, especially bacteria, were selectively introduced on the commercial level. During World War I, industrial microbiology was directed at the production of chemicals, especially glycerol, acetone, and butanol. Techniques were developed gradually for pure-culture fermentation and strain improvement, but the major advance in industrial microbiology occurred during World War II with the large-scale production of penicillin by submerged-culture fermentation.

In the 1950s, at a time when microbial processes for chemicals production were being replaced by petrochemical processes, industrial microbiology shifted its focus to the production of therapeutic agents, especially antibiotics. Advances in molecular biology have greatly increased the potential applications of industrial microbiology in areas such as therapeutics, diagnostics, environmental protection, and agriculture. The techniques of genetic engineering, along with technology developments in bioprocessing, make possible large-scale production of complex natural compounds that would otherwise be very difficult to obtain.

Strain development. With the exception of the food industry, few commercial fermentation processes use wild strains of microorganisms isolated directly from nature. Of the many thousands of microbial species, few are used commercially, and fewer still are used as hosts for genetically engineering genes. Process development occurs in large part by strain improvement directed at increasing product yield, enhancing growth on cheaper substrates, and simplifying purification.

Strain development is achieved by either a traditional mutation and selection program or direct genetic manipulation. The dominant reason for strain improvement by mutation and selection is the alteration of the regulatory mechanisms of an organism so that maximum metabolic energy is directed to a single product, either a protein or a product from a metabolic pathway. The approach, which is relatively rapid, is inexpensive, and has shown good results, involves successive rounds of mutagenesis and screening for the desired mutants. However, it is limited by the initial genetic potential of the organism.

The recombinant deoxyribonucleic acid (DNA) approach has succeeded in introducing new genetic material into a convenient host microorganism and amplifying genetic material. About 20% of the synthesizing capacity of a bacterium can be devoted to a single polypeptide or protein. *See* RECOMBINATION (GENETICS).

Fermentation. Commercial microbial compounds are produced in two distinct phases: fermentation and product recovery. Fermentation procedures are developed by determining optimal conditions for product synthesis. Commercial production usually occurs in a batch fermentor, where gas of controlled composition and flow is bubbled through a stirred pure microbial culture suspended in a liquid medium of optimum nutrient composition. The maintenance of sterile conditions during the fermentation is crucial, and so all liquids and gases are sterilized before they are fed to the fermentor. Key operating variables such as pH, temperature, and dissolved oxygen content are controlled at their optimal values during fermentation for rapid cell growth and product formation.

Product recovery and purification involves a series of operations. The first steps usually involve cell disruption or the separation of the cell or cellular debris from the fluid medium, typically through centrifugation and filtration. Later stages of purification include finer membrane filtration, extraction, precipitation, and chromatography. These techniques must be optimized with respect to the specific biochemical and physical differences between product and impurities. *See* CENTRIFUGATION; FERMENTATION; FILTRATION; FOOD FERMENTATION; STERILIZATION.

Applications. The production of certain foods and beverages was an early application of industrial microbiology. Now, the fields of mineral recovery, medicine, environmental protection, and food and agriculture are using similar techniques.

Mineral recovery. Bacterial leaching reactions have been used to alter metal-bearing minerals, usually converting them to more soluble forms before the metals are extracted. Such operations can result in improved extraction rates in comparison to those of conventional processes, which are usually conducted on ore waste dumps and heaps. Large-scale commercial applications of biochemical mining and extraction have been limited mainly to copper and uranium. Through genetic engineering techniques, microorganisms are being developed that are more compatible with high concentrations of metal ions, low levels of oxygen, and high temperatures. For some applications, a mixture of several strains is the best, because some are more resistant to high temperature, whereas others tolerate higher concentrations of heavy metals. The most common strain, involved in copper and iron sulfide leaching, is *Thiobacillus ferrooxidans*.

Bacteria derive all their metabolic needs from inorganic substances, such as carbon dioxide and metal sulfides, and organic compounds, such as carbon dioxide, sugars, and organic acids. Certain combinations of these organics can enhance the effectiveness of the bacteria, as can nitrogen-enriched leachates and surfactants. In metallurgical leaching reactions, bacteria aid the oxidation of complex metal sulfides to form water-soluble metal sulfates. In addition, bacterial leaching has been used to extract other metal values from sulfide ores: zinc from sphalerite, lead from galena, and cobalt and nickel from base-metal sulfide ores. Nonmetal ores have also been processed with that method, particularly coal, which often contains high concentrations of sulfur that cause air pollution when burned.

Besides enhancing or inhibiting the recovery of metal values from ores, bacteria are being used to precipitate or accumulate metal. The process, known as bioaccumulation, normally involves the adsorption of metal ions on the bacteria, which are then chemically transformed to an insoluble precipitate. *See* METAL AND MINERAL PROCESSING.

Medical applications. Diagnostic applications have developed rapidly, beginning with the introduction of diagnostic tests that use monoclonal antibodies. A monoclonal antibody is produced by a hybridoma, a cell formed by the fusion of an antibody-producing cell with a cancerous myeloma cell. Monoclonal antibodies specifically combine with their complementary antigen and form the diagnostic basis for many drug-monitoring tests, syphilis and rubella testing, blood testing, cancer screening, and pregnancy tests. *See* MONOCLONAL ANTIBODIES.

The most visible products of industrial microbiology are therapeutics for human health. Microbial synthesis is the preferred method of production for most health care drugs with complex chemistry. Microorganisms still have a remarkable ability for producing new commercial antibiotics, the largest class of drugs, and for continued yield improvement. With recombinant DNA technology, many proteins and polypeptides that are present naturally in the human body in trace amounts can be produced in large amounts during fermentation of recombinant microorganisms. Full regulatory approval for new biotechnology products usually takes 8 to 9 years from initial synthesis, but already the following products have been approved: recombinant insulin for the treatment of diabetes; recombinant factor VIII for hemophilia; somatropin human growth hormone for growth deficiencies in children; tissue plasminogen activator for acute myocardial infarction; interferon for hairy cell leukemia; a vaccine for hepatitis B; and a monoclonal antibody to counteract rejection of transplanted kidneys. Other areas of possible application are immunomodulators, including interferons, colony stimulating factors, and interleukins; blood proteins, including anticoagulants and thrombolytic agents; and monoclonal antibodies with specific anticancer activity. *See* ANTIBIOTIC; INSULIN.

Pharmaceutical agents produced from recombinant microorganisms must meet stringent purity criteria. These products must be completely free of microbial cells, have no more than 10 picograms of DNA contamination per human dose, and be free of viral contamination. Recombinant products are usually proteins that must be separated from other contaminating proteins through several different and complementary methods of purification. Licenses issued for their production are generally process-specific, in contrast to more simple antibiotics that can be precisely defined chemically.

Environmental applications. Microorganisms have important roles in the natural cycling of elements because of the wide range of microbial degradation activities. Microbial activities have long been the basis for sewage treatment facilities, and industrial and hazardous waste cleanup, or bioremediation, has become important. The usual first step in bioremediation is the isolation of microorganisms, often from an environment contaminated by the toxic compounds, that are capable of growing on and degrading a specific toxic chemical waste. Environmental control during bioremediation of wastewater or contaminated soil is crucial because proper nutrient supply, aeration, and mixing can greatly enhance degradation rates. *See* HAZARDOUS WASTE; SEWAGE.

Many recalcitrant synthetic compounds are released into the environment. Some of these persist in nature because of their low rates of biodegradation by indigenous microorganisms. Bioremediation successes have been achieved by using native bacteria to degrade petroleum products (such as kerosine, benzene, and toluene that have leaked from underground storage tanks), toxic chlorinated herbicides (such as 2,4-dichlorophenoxyacetic acid [2,4-D]), and toxic biocides (such as pentachlorophenol).

Food and agricultural applications. Some of the oldest and most established areas of industrial microbiology concern food and beverage products, such as the production and use of brewer's yeast and baker's yeast. Food and agricultural applications are more limited by cost constraints than are medical applications, with the cost of raw materials often the major economic factor. The beer industry, which produces the highest-value product of all the biotechnology industries, has developed cost-containing strategies such as spent mash and carbon dioxide recovery, efficient flocculation and clarification, nonaseptic fermentation, and continuous large-scale fermentation. *See* CHEESE; DISTILLED SPIRITS; MALT BEVERAGE; WINE.

The food industry and the detergent industry are the major users of industrial enzymes produced by microbial fermentation. Detergents, especially in Europe, often contain protein-degrading enzymes (proteases). In the food industry, amylases convert starch to glucose, and glucose isomerase converts glucose to fructose. *See* DETERGENT; FOOD MANUFACTURING.

Agricultural applications use technology similar to that of medical applications. There are many agricultural applications for genetic manipulation of plants and possibly animals. Plant cultures have been applied to the production of important natural compounds and the improvement and propagation of crop plants. Because the production of natural compounds by plant cells must compete with less complex extraction from whole plants, the only commercial plant-cell culture process thus far appears to be the production of the red dye shikonin in Japan. Mass propagation by tissue culture and improvement of plants have received more commercial attention. *See* BIOTECHNOLOGY; BREEDING (PLANT); GENETIC ENGINEERING; PETROLEUM MICROBIOLOGY. Roger Korus

Bibliography. E. Antébi and D. Fishlock, *Biotechnology Strategies for Life*, 1986; C. Brown, I. Campbell, and F. Priest, *Introduction to Biotechnology*, 1987; A. L. Demain and N. A. Solomon (eds.), *Biology of Industrial Microorganisms*, 1985; H. L. Ehrlich and

C. L. Brierley, *Microbial Mineral Recovery*, 1990; R. Hack (ed.), *The International Biotechnology Handbook*, 1989; D. P. Labeda, *Isolation of Biotechnological Organisms from Nature*, 1990; P. Prave et al. (eds.), *Fundamentals of Biotechnology*, 1987; I. K. Vasil (ed.), *Biotechnology: Perspectives, Policies and Issues*, 1987.

Industrial trucks

Manually propelled (handtrucks) or powered carriers for transporting materials over level, slightly inclined, or slightly declined running surfaces. Some industrial trucks can lift and lower their loads, and others can also tier them. In any event, all such trucks maintain contact with the running surface over which they operate and, except when towed by a chain conveyor, follow variable paths of travel as distinct from conveying machines or monorails.

In industry the principle of handling materials in unit loads has developed in parallel with the increased use of powered industrial trucks, particularly the forklift type. These mobile mechanical handling aids have removed the limitations that existed when the weight and size of a load for movement and stacking depended mainly on the ability of a person to lift it manually. The unit-load principle of materials handling underlies the skid-platform and the pallet-forklift methods of operation. Both methods, especially the latter, have revolutionized handling techniques and equipment and even production equipment. *See* MATERIALS HANDLING.

Arthur M. Perrin

Bibliography. Icon Group International, Inc., *The 2006–2011 World Outlook for Self-Propelled, Electric, Gasoline and Other Power Work Trucks, Fork Lifts, and Tractors Fitted or Not Fitted with Lifting and Handling Equipment*, 2005; D. E. Mulcahy, *Materials Handling Handbook*, 1998.

Industrial wastewater treatment

A group of unit processes designed to separate, modify, remove, and destroy undesirable substances carried by wastewater from industrial sources. United States governmental regulations have been issued for volatile organic substances, designated priority pollutants, aquatic toxicity as defined by a bioassay, and in some cases nitrogen and phosphorus. As a result, sophisticated technology and process controls have been developed for treating industrial wastewater. It is also necessary to implement better water management and waste minimization practices to reduce waste loads and operating cost of treatment facilities.

Wastewater characterization. Wastewater must be characterized to determine if it will meet regulatory requirements. Characterization is also necessary to collect the information needed to design processes that can effectively reduce and treat the wastewater. Before choosing the treatment technologies, some key points must be considered, including understanding the production processes and how wastewater is generated, measuring the wastewater volumes and flow rates, and sampling and analyzing the physical, chemical, and biological characteristics.

Water-pinch technology emerged as a significant new approach to the problems of industrial water reuse, wastewater minimization, and effluent treatment system design. This technology allows chemical, civil, and environmental engineers to analyze water-using processes before design and operation, as well as after, to minimize freshwater consumption and wastewater generation.

Zero emission, zero discharge, cleaner production, waste minimization, pollution prevention, design for environment, material substitution, and dematerialization are relevant terms. The U.S. Environmental Protection Agency (EPA), the United Nations Industrial Development Organization (UNIDO), and other national and international organizations have promoted these terms.

Source treatment. Wastewater streams that are toxic or resistant to conventional treatment methods (refractory) should be treated at the source, and there are a number of source treatment technologies available. Metal precipitation enables the removal of bioinhibitors or biotoxicants. Chemical oxidation removes bioinhibitors and improves biodegradation rates. Wet air oxidation, which oxidizes organic materials at high temperature and pressure [14 kilopascals (2000 lb/in.^{2}) and 288°C (550°F)], is restricted to very high concentrations of these substances. Macroreticular (macroporous) resins are specific for removing particular organic materials, and the resin is regenerated and reused. Membrane processes, particularly reverse osmosis, are high-pressure operations in which water passes through a semipermeable membrane, leaving the contaminants in a concentrate. Other source treatment processes include stripping, filtration, and carbon adsorption. *See* HAZARDOUS WASTE; MEMBRANE SEPARATIONS.

Pretreatment and primary treatment. These processes address the problems of equalization, neutralization, removal of oil and grease, removal of suspended solids, and precipitation of heavy metals (**Fig. 1**).

Equalization. Effective operation of a wastewater treatment plan requires controlling the flow and loading to the facility. Equalization is used to achieve this objective. If the wastewater flow rate is relatively constant, as in a pulp and paper mill, a constant-volume basin is used in which the concentration is equalized. If both flow and concentration change, a variable-volume basin is used in which the inflow is variable and the outflow is constant. In this way, the mass load is equalized. Many industries are subject to spills and accidental discharges, for example, black liquor spills in the pulp and paper industry. In these cases, a spill basin is provided to retain spills and discharge them at a controlled rate to the equalization basin.

Neutralization. Wastewater permits usually require a pH in the range of 6 to 9. Biological treatment

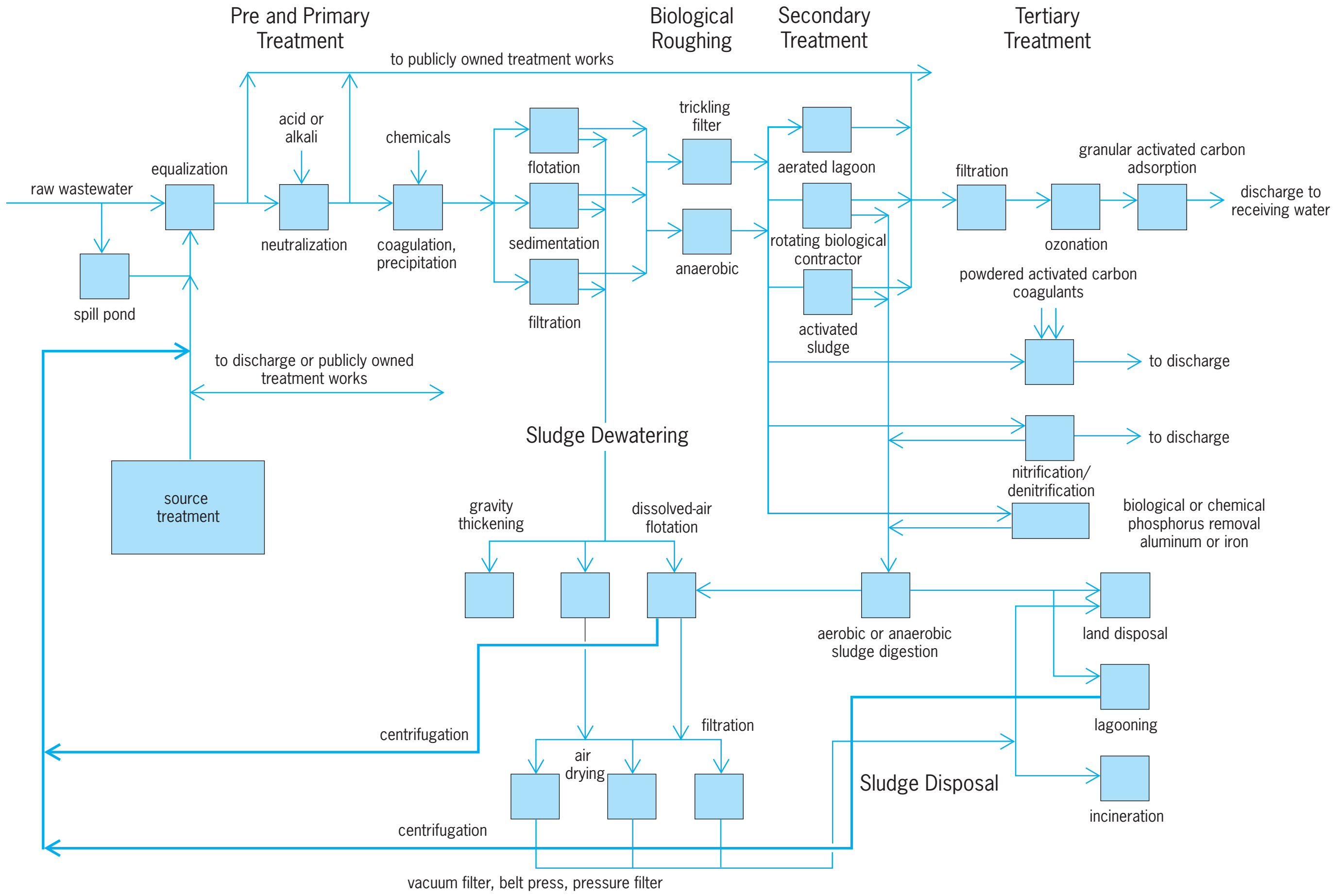


Fig. 1. Flowchart showing alternative wastewater treatment technologies. Flow returns are shown by heavy lines.

requires a pH of 6.5 to 8.5, and chemical coagulation processes have defined pH ranges. For this reason, acidic or alkaline wastewater streams require neutralization. Acidic wastewaters can be neutralized with lime ($CaOH_2$), caustic (NaOH), or magnesium hydroxide [$Mg(OH)_2$]. Magnesium hydroxide frequently has the advantage of producing less but more concentrated sludge. Since pH is a logarithmic function, highly acidic wastewaters require a two-stage neutralization process to maintain an effluent with a stable pH. Where the wastewater pH is fairly constant, a limestone ($CaCO_3$) bed can be used. Alkaline wastewaters are neutralized with sulfuric acid (H_2SO_4) or flue gas (CO_2), which forms carbonic acid when dissolved in water. *See* PH.

Oil, grease, and suspended solids. Free oil is removed by gravity separation. Large wastewater volumes containing oil, such as those issuing from a petroleum refinery, use an American Petroleum Institute (API) separator. The separator is a long, rectangular shallow basin in which the lower-specific-gravity oil rises to the surface and is skimmed off. Effluent oil concentrations are about 50 mg/liter. Lower-volume wastewater is treated in a corrugated plate separator that contains closely spaced plates, permitting a higher efficiency of oil removal. Effluent oil concentrations of 10 mg/liter are possible. In the case of emulsified oil, the emulsion must be broken using polymers or surfactants, enabling the creation of oily aggregates (flocs) that can be separated by physical means. *See* EMULSION; SURFACTANT.

Suspended solids can be removed by sedimentation or flotation. The efficiency of suspended-solids removal by sedimentation is related to the overflow rate [m^3/m^2 (gal/ft^2)], and for most wastewaters removal efficiencies will range 65–90%. Flotation is used to remove low concentrations of oil, grease, and nonsettleable solids. In dissolved air flotation, a portion of the effluent is pressurized in the presence of excess air, and the additional air goes into solution. When the pressurized liquid is mixed with the influent wastewater in the flotation unit at near-atmospheric pressure, the dissolved air comes out of solution in the form of small bubbles that attach themselves to and become enmeshed in globules and particles that rise to the surface and are skimmed off. Induced air flotation serves the same purpose, but it generates air bubbles through rotary inductors. *See* SEDIMENTATION (INDUSTRY).

Colloidal particles, turbidity, and color can be removed by coagulation. In this process, a coagulant, usually alum [$Al_2(SO_4)_3$] or iron(III) chloride [$FeCl_3$], is added to the wastewater. The positive aluminum and iron ions neutralize the negatively charged particles and generate a floc that is subsequently removed in a settling basin. *See* COLLOID; PARTICULATES.

Heavy-metal precipitation. Heavy metals are conventionally removed by precipitation. Most metals will precipitate as the hydroxide, carbonate, or sulfide. Carbonate and hydroxide precipitates are highly pH-dependent, according to the metal. Therefore, close pH control is essential. The sulfide is less pH-dependent. But because sulfide in the form of hydrogen sulfide (H_2S) presents an odor nuisance and a health hazard, sulfide precipitation is mainly used as a polishing step. In co-precipitation, heavy-metal ions are adsorbed onto an aluminum hydroxide or iron hydroxide floc. While this process is highly effective for removing most heavy metals, a large amount of sludge is generated that requires disposal. In the treatment of inorganic wastewater, emphasis is placed on metal recovery using techniques such as ion exchange, reverse osmosis, and electrochemical processes. A number of technologies are available for removing various metals from the wastewater stream. *See* ELECTROCHEMICAL PROCESS; ION EXCHANGE; PRECIPITATION (CHEMISTRY).

Biological treatment. Aerobic biological treatment is used to remove biodegradable organics. Depending on the effluent requirements, the nature of the wastewater, and the availability of land, several process options are available. In the activated-sludge process, the biomass generated is settled in a clarifier and recycled back to the aeration basin to build a high concentration of biomass. The process performance is related to the organic loading. Effluent concentrations of biodegradable organics in the order of 25 mg/liter are readily achievable. To maintain good sludge quality, a plug-flow configuration or batch treatment should be used in the case of readily degradable wastewaters, and a complete mix configuration for the more refractory wastewaters. Excess sludge generated in the process is disposed of from the final clarifier. Inhibition and poor settleability of the biosolids are common disturbances reported in cases of biological treatment of wastewater from petrochemical, pharmaceutical, paper, and other industries. In some cases, membrane separation technologies are used instead of sedimentation, and these systems are known as membrane biological reactors (MBR).

Where large land areas are available, an aerated lagoon system is applicable for treating nontoxic wastewaters, such as generated by pulp and paper mills. Three basins in series are used. The first basin is completely mixed in order to maximize biodegradation. In the second basin, at reduced power level, solids are permitted to settle, and then undergo anaerobic degradation and stabilization. A third settling basin is usually employed to reduce effluent suspended solids. The advantages of an aerated lagoon are simplicity of operation and the need to remove the sludge only about every 10 years.

Fixed-film processes include the trickling filter and the rotating biological contactor. In these processes, a biofilm is generated on a surface, usually plastic. As the wastewater passes over the film, organics diffuse into the film and are biodegraded. The trickling filter is primarily used as a pretreatment process for high concentrations of biodegradable organics in readily degradable wastewaters. Removal efficiencies of about 50% are usually achieved. The rotating biological contactor has been successfully used for treating wastewaters with low-to-moderate concentrations of biodegradable organics, such as generated by petroleum refineries and paper mills. *See* BIOFILM.

Anaerobic processes are sometimes used before aerobic processes for treating high-strength, readily degradable wastewaters. The primary advantages of the anaerobic process are low sludge production and the generation of energy in the form of methane (CH_4) gas. In the anaerobic process, organic materials are broken down to volatile organic acids, primarily acetic acid, which in turn is broken down to methane and carbon dioxide (CO_2). High-rate processes include the anaerobic filter, in which the anaerobic organisms are grown on a support medium; the anaerobic contact process, which is similar to the activated-sludge process and consists of a reactor, a clarifier, and a sludge recycle; and the upflow anaerobic sludge bed, which generates a granular sludge through which the wastewater is passed (**Fig. 2**). Where land area is available, low-rate processes have the advantage of greater process stability against spilling high concentrations of organics or accidentally dumping wastewater. *See* BIODEGRADATION; SEWAGE DISPOSAL; SEWAGE TREATMENT.

Physical-chemical treatment. Biological processes can remove only degradable organics. Nondegradable organics can be present in the influent wastewater or generated as oxidation by-products in the biological process. Many of these organics are toxic to aquatic life and must be removed before discharging the effluent. The most common technology used is adsorption on activated carbon. Granular carbon filters can be used as a posttreatment in which the refractory organics are adsorbed on the carbon. The granular carbon is regenerated for reuse. Granular media filtration is usually employed prior to the carbon columns. Alternatively, powdered activated carbon can be used in conjunction with the activated-sludge process. In this case, the powdered activated carbon is added prior to or into the activated-sludge aeration basin. The biomass degrades the degradable organics, and the carbon adsorbs the nondegradable organics. The carbon-biosludge can be regenerated or disposed of in a landfill. *See* ACTIVATED CARBON; ADSORPTION.

In some cases, toxic and refractory organics can be pretreated by chemical oxidation using ozone, catalyzed hydrogen peroxide, or advanced oxidation processes. In this case, the objective is not mineralization of the organics but detoxification and enhanced biodegradability.

Biological nitrogen removal, including both nitrification and denitrification, is used to remove ammonia from wastewaters. While this process is predictable in the case of municipal wastewaters, many industrial wastewaters inhibit the nitrifying organisms. For example, while nitrification will occur at

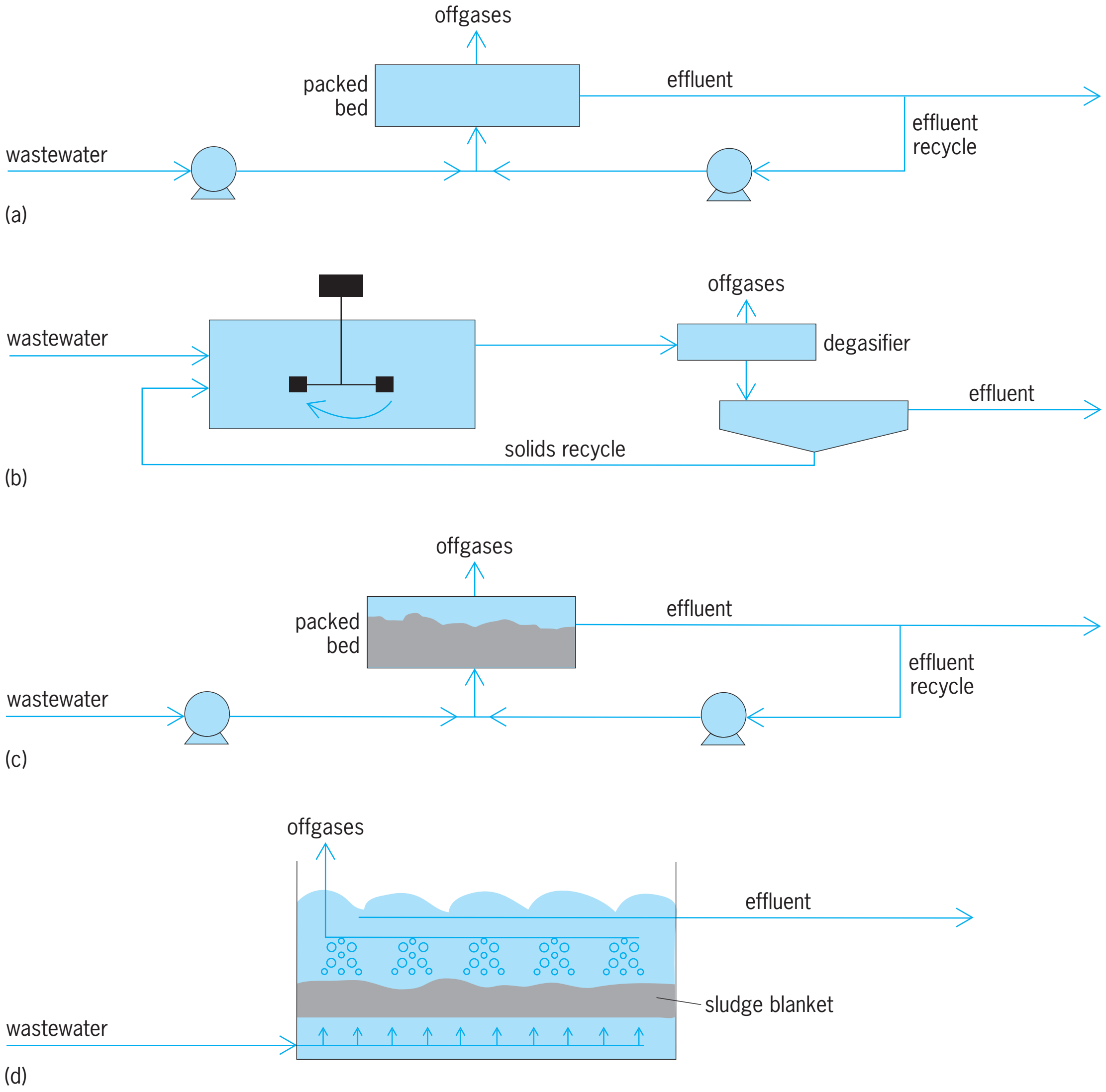


Fig. 2. Anaerobic wastewater treatment processes. (*a*) Anaerobic filter reactor. (*b*) Anaerobic contact reactor. (*c*) Fluidized-bed reactor. (*d*) Upflow anaerobic-sludge blanket.

a sludge age of 3 days at 20°C (68°F) in municipal wastewater, as much as 25 days may be required in the case of wastewater from a chemical plant. Sludge age is defined as the average length of time that the organism is under aeration. Inhibited conditions slow the growth rate of the organisms, and the sludge age must be increased. This step, in turn, increases the required aeration basin volume. In extreme cases of toxicity to the nitrifiers, powdered activated carbon has been required to adsorb the toxic molecules to permit nitrification to proceed. *See* BACTERIA.

Volatile organics can be removed by air or steam stripping. Air stripping is achieved by using packed or tray towers in which air and water counterflow through the tower. In cases where air stripping is used, the offgas containing the volatiles frequently must be treated by using vapor-phase carbon adsorption or a combustion chamber. In steam stripping, the liquid effluent from the column is separated as an azeotropic mixture. *See* AZEOTROPIC MIXTURE; STRIPPING.

Residuals management. Residuals are defined as any waste that is solid, liquid, or gas. Wastewater treatment operations usually produce sludge and solids. Disposal of those materials can be a major consideration when it comes to selecting wastewater unit processes. Every plant should perform residuals inventory, which should go back to the source of the residuals. A residuals management plan can be formulated to find the most cost-effective method for disposing of these wastes.

Sludge mixtures frequently contain less than 5% solids and therefore need to be thickened and dewatered for cost-effective disposal. Waste-activated sludge is usually stabilized by anaerobic digestion or by aerobic digestion, in which the degradable solids are oxidized by prolonged aeration. Ultimate disposal can be on land, by lagooning, or by incinerating the dewatered material. Return flows from thickening and dewatering are returned to the plant influent. It is frequently cost-effective to dry or evaporate sludges to reduce the volume and increase heat content in case of incineration.

It is important to separate hazardous from nonhazardous sludges to minimize disposal costs. There are two types of landfills for industrial waste-treatment residuals: those that accept only nonhazardous waste and those that accept hazardous waste. Landfills are designed to prevent the contamination of ground water and the migration of contaminants from the landfill. Therefore, landfills must meet very stringent technical requirements. *See* SEWAGE SOLIDS.

Landfill leachates. Most landfill leachates have high and variable concentrations of organic and inorganic substances. All municipal and most industrial landfill leachates are amenable to biological treatment and can be treated anaerobically or aerobically, depending on the effluent quality desired. Activated carbon has been used to remove nondegradable organics. In Europe, some plants use reverse osmosis to produce a high-quality effluent. *See* WATER POLLUTION.

W. Wesley Eckenfelder; Noah I. Galil

Bibliography. J. B. Blackburn, *The Industrial Wastewater Systems Handbook*, 1998; W. W. Eckenfelder, *Industrial Water Pollution Control*, 3d ed., 1999; W. W. Eckenfelder, *Principles of Water Quality Management*, 1980, reprint 1991; J. D. Edwards, *Industrial Wastewater Treatment*, 1995; N. Galil, M. Rebhun, and Y. Brayer, Disturbances and inhibition in biological treatment of an integrated oil refinery, *Water Sci. Technol.*, 20(10):21–29, 1988; P. W. Lankford and W. W. Eckenfelder, *Toxicity Reduction in Industrial Effluents*, 1990; J. G. Mann and Y. A. Liu, *Industrial Water Reuse and Wastewater Minimization*, 1999; Metcalf & Eddy, Inc., G. Tchobanoglous et al., *Wastewater Engineering: Treatment, and Reuse*, 4th ed., 2003; R. K. Miller and M. E. Nemerow, *Industrial Wastewater Treatment*, 1991; J. W. Patterson, *Industrial Wastewater Treatment Technology*, 1985; L. K. Wang et al., *Handbook of Industrial and Hazardous Waste Treatment*, 2004; F. Woodard, *Industrial Waste Treatment Handbook*, 2001.

Inert gases

The inert gases, listed in the **table**, constitute group 0 of the periodic table of the elements. They are now better known as the noble gases, since stable compounds of xenon have been prepared. The term rare gases is a misnomer, since argon is plentiful and helium is not rare in the United States and some other countries.

The inert gases

Name	Symbol	Atomic number	Atomic weight
Helium	He	2	4.0026
Neon	Ne	10	20.183
Argon	Ar	18	39.948
Krypton	Kr	36	83.80
Xenon	Xe	54	131.30
Radon	Rn	86	(222)

All these gases occur to some extent in the Earth's atmosphere, but the concentrations of all but argon are exceedingly low. Argon is plentiful, constituting almost 1% of the air.

All isotopes of radon are radioactive, the longest lived having a half-life of about 4 days. Each of the other inert gases has at least two stable (nonradioactive) isotopes, in addition to one or more radioactive isotopes.

All the gases are colorless, odorless, and tasteless. They are all slightly soluble in water, the solubility increasing with increasing molecular weight. They can be liquefied at low temperatures, the boiling point being proportional to the atomic weight. All but helium can be solidified by reducing the temperature sufficiently, and helium can be solidified at temperatures of 0 to 1 K (−459.67 to −457.87°F) by the application of an external pressure of 25 atm (2.5×10^3 kilopascals) or more.

The noble gases are all monatomic. The outer shell of each of the atoms (unless strongly excited by radiation, electron bombardment, or other disturbing effects) is completely filled with electrons, and the gases are generally chemically inert. However, since the outer electrons in xenon are relatively far from the nucleus, they are held rather loosely; and they are capable of interacting with the outer electrons of fluorine atoms and the atoms of some other elements to form fairly stable compounds, such as XeF_2 and XeF_4. Radon can presumably form compounds similar to those of xenon. A fluoride of krypton which is stable at −8°C (17.6°F) has been prepared. No compounds of any of the lighter noble gases have been obtained. *See* XENON.

In an electric discharge, as in a mass spectrometer, very short-lived ions and molecules of all the noble-gas atoms can be formed; examples are Ar_2^+, HgHe, HgNe, HgAr, $(ArKr)^+$, and $(NeNe)^+$. In the presence of methane, $(XeCH_3)^+$ and other highly unstable ions can be formed.

All the inert gases except helium and radon are produced in concentrated form by the liquefaction and distillation of air, followed by special purification processes. Helium is obtained from certain natural gases containing 0.5% or more He. Radon is obtained by collecting the gas, called radium emanation, given off in the radioactive decay of radium. *See* AIR SEPARATION; ARGON; HELIUM; KRYPTON; NEON; PERIODIC TABLE; RADON; XENON.

A. W. Francis

Bibliography. V. G. Fastovskii, *Inert Gases*, 1967; B. L. Smith and J. P. Webb, *The Inert Gases: Model Systems for Science*, 1971.

Inertia

That property of matter which manifests itself as a resistance to any change in the motion of a body. Thus when no external force is acting, a body at rest remains at rest and a body in motion continues moving in a straight line with a uniform speed (Newton's first law of motion). The mass of a body is a measure of its inertia. *See* MASS.

Leo Nedelsky

Inertia welding

A welding process used to join similar and dissimilar materials at very rapid speed. It is, therefore, a very attractive welding process in mass production of good-quality welds. The ability to join dissimilar materials provides further flexibility in the design of mechanical components.

Principles. Inertia welding is a type of friction welding which utilizes the frictional heat generated at the rubbing surfaces to raise the temperature to a degree that the two parts can be forged together to form a solid bond. The energy required for inertia welding comes from a rotating flywheel system built into the machine. **Figure 1** illustrates the principle of the process. Like an engine lathe, the inertia welding machine has a headstock and a tailstock. One workpiece held in the spindle chuck (usually with an attached flywheel) is accelerated rapidly, while the other is clamped in a stationary holding device of the tailstock. When a predetermined spindle speed is reached, the drive power is cut and the nonrotating part on the tailstock is pushed against the rotating part under high pressure. Friction between the rubbing surfaces quickly brings the spindle to a stop. At the same time the stored kinetic energy in the flywheel is converted into frictional heat which raises the temperature at the interface high enough to forge the two parts together without melting.

During the process the material near the interface is softened and expelled from the original contacting surfaces to form a flash. This brings the nascent sublayers of the material close together to form a metallurgical bond. Therefore, the welding process is autogenous and carried out in solid state without the need of flux or filler material.

Weld characteristics. Inertia welds of similar materials consist of a thin layer of a severely deformed but highly refined grain structure at the weld interface. The heat-affected zone is very narrow as compared with any fusion welding process. For a properly welded joint, the weld strength is significantly higher than that of the base material. This is because during inertia welding the forging process improves the hardness of the material at the weld zone by means of work hardening. An inertia weld of carbon steels can have an average hardness at the weld some 25% higher than that of the base material. A tensile test will result in a fracture occurring at the base material instead of at the weld. If the specimen is annealed after welding, the interface would completely disappear as if the materials were not welded.

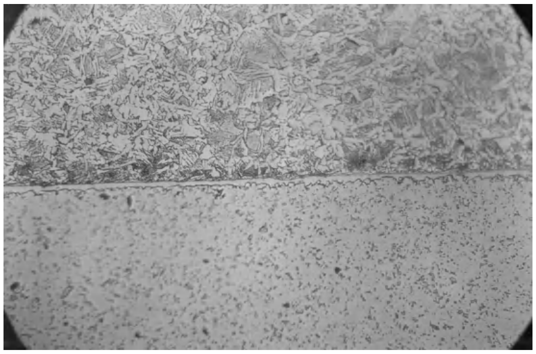


Fig. 2. Micrograph of a joint of SAE 1020 steel welded to tungsten carbide.

One of the principal advantages of inertia welding is its ability to join dissimilar metals which would be unweldable or uneconomical to join by any other means. **Figure 2** is a micrograph of a specimen of SAE 1020 steel welded to tungsten carbide. The process is also used to join aluminum to steel and stainless steel, aluminum to copper, copper to steel, copper to titanium, and zirconium alloys to each other. The weldability chart (**Fig. 3**) shows that some metal combinations can be welded to achieve strength at the weld equal to or better than the strength of the weaker material in the pair. In some cases, however, it may be necessary to perform an appropriate postweld heat treatment to realize the full weld strength. The chart indicates that there are other combinations that can be welded, but with strengths usually less than the strength of the weaker material.

The mechanisms which cause bonding of dissimilar metals are rather complex. Factors such as physical properties, surface energy, crystalline structure, mutual solubility, and possible formation of intermetallic compounds from the materials being welded may all play a role in the bonding mechanism. Since the key to achieving a metallurgical bond in solid state is to break up and remove the contaminated surface layers, metals with a substantial difference in hardness are generally more difficult to weld. In such cases, the hard material incurs practically no deformation, while the soft one has already been severely deformed or extruded. In

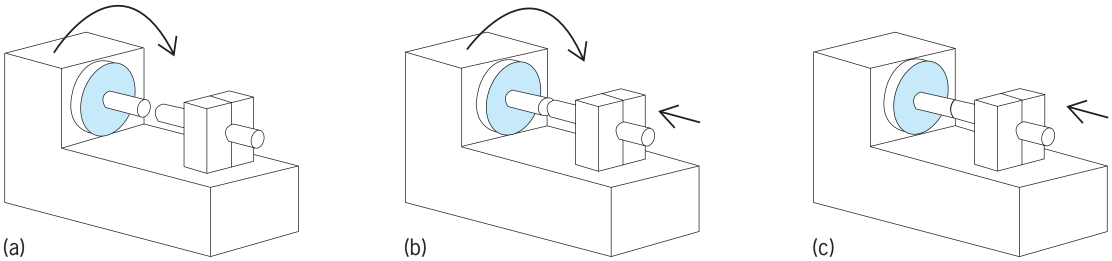
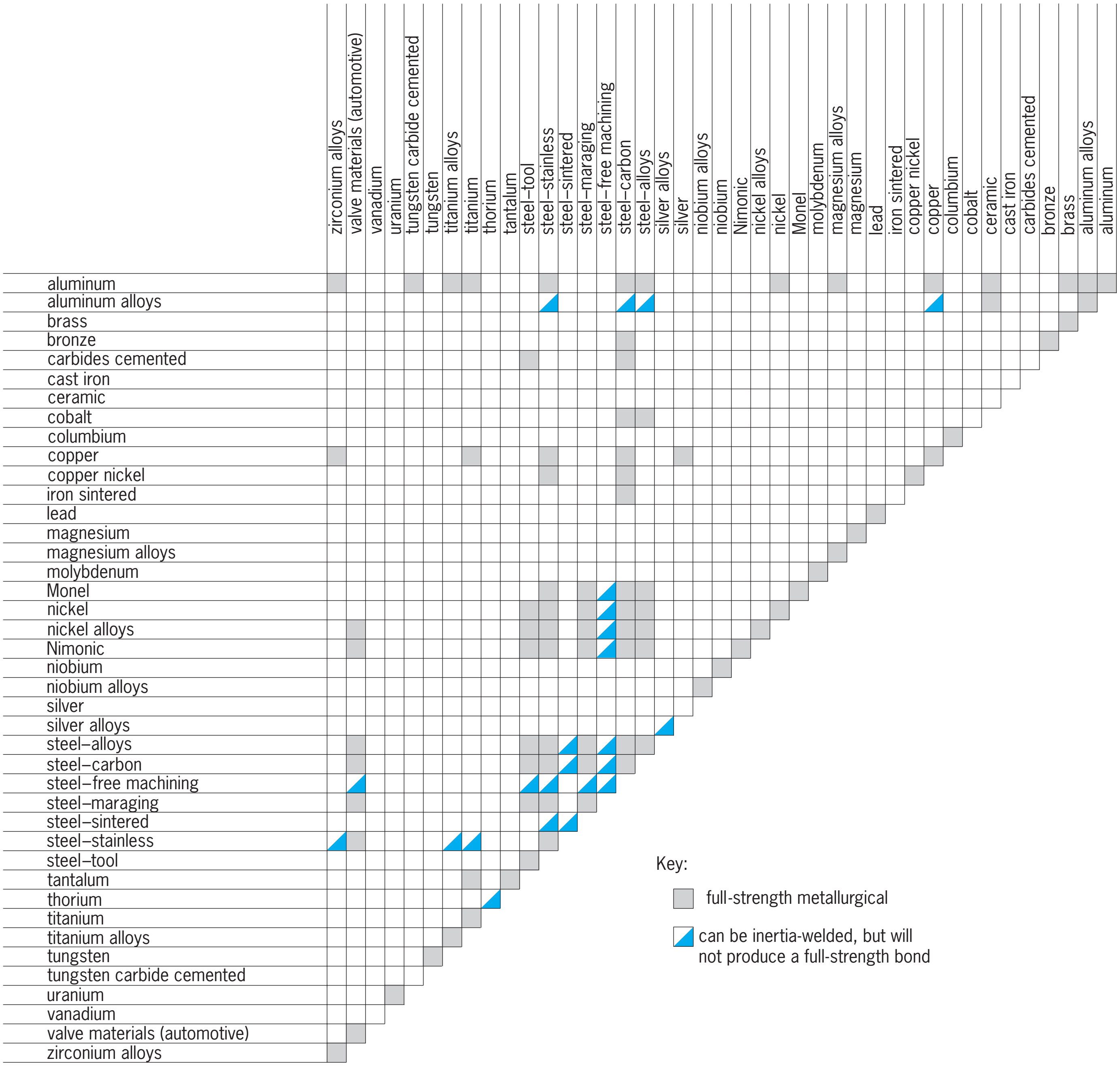


Fig. 1. Work sequence of inertia welding process. (*a*) Acceleration of flywheel and chuck-holding workpiece to preset speed. (*b*) At speed, drive is cut and workpieces are thrust together. (*c*) Workpieces bond together to stop flywheel rotation; weld is complete.



	zirconium alloys	valve materials (automotive)	vanadium	uranium	tungsten carbide cemented	tungsten	titanium alloys	titanium	thorium	tantalum	steel–tool	steel–stainless	steel–sintered	steel–maraging	steel–free machining	steel–carbon	steel–alloys	silver alloys	silver	niobium alloys	niobium	Nimonic	nickel alloys	nickel	Monel	molybdenum	magnesium alloys	magnesium	lead	iron sintered	copper nickel	copper	columbium	cobalt	ceramic	cast iron	carbides cemented	bronze	brass	aluminum alloys	aluminum
aluminum	■				■		■	■				■				■	■							■			■					■			■				■	■	■
aluminum alloys												◪				◪	◪															◪			■					■	
brass																																							■		
bronze																■																						■			
carbides cemented											■					■																									
cast iron																																									
ceramic																																									
cobalt																■	■																								
columbium																																	■								
copper	■							■				■				■			■													■									
copper nickel												■				■															■										
iron sintered																■																									
lead																													■												
magnesium																												■													
magnesium alloys																											■														
molybdenum																										■															
Monel												■		■	◪	■	■								■																
nickel											■	■		■	◪	■	■							■																	
nickel alloys		■									■	■		■	◪	■	■						■																		
Nimonic		■									■	■		■	◪	■	■					■																			
niobium																					■																				
niobium alloys																				■																					
silver																																									
silver alloys																		◪																							
steel–alloys		■									■	■	◪	■	◪	■	■																								
steel–carbon		■									■	■	◪	■	◪	■																									
steel–free machining		◪									◪	◪		◪	◪																										
steel–maraging		■									■	■		■																											
steel–sintered												◪	◪																												
steel–stainless	◪	■					◪	◪				■																													
steel–tool											■																														
tantalum								■		■																															
thorium									◪																																
titanium								■																																	
titanium alloys							■																																		
tungsten						■																																			
tungsten carbide cemented																																									
uranium				■																																					
vanadium																																									
valve materials (automotive)		■																																							
zirconium alloys	■																																								

Key:

■ full-strength metallurgical

◪ can be inertia-welded, but will not produce a full-strength bond

Fig. 3. Weldability between metals by inertia welding.

general, inertia welding can be applied with steels, aluminum, titanium, cobalt, nickel, and their alloys, copper, brass, and some bronzes, refractory materials, precipitation-hardened metal alloys, sintered materials, and heat-treated materials. The process cannot be successfully applied to cast irons, dry bearing materials, beryllium, and other materials where a distinct weak phase is present in the microstructure, such as in graphite, manganese sulfide, or there is a large amount of free lead or tellurium.

Applications. In addition to the advantages of high weld strength and short welding-cycle time, inertia welding machines can be readily automated for mass production. The automotive and truck industry is the major user of the process. Typical applications are engine valves (a medium-carbon-steel bar stem welded to a forged heat-resistant-alloy steel head), base-cup shock absorber assemblies, universal joints, axle housings, and hydraulic cylinder components. Auto makers, as well as other manufacturers, are constantly looking for ways to reduce the weight of components without sacrificing strength. An inexpensive way of achieving this goal is to replace standard open-die forgings or castings by weldments of precision closed-die forgings or castings welded to less expensive tubes, bar stocks, or stampings. Inertia welding of a tool pin joint to a drill pipe is one typical application in the oil drill industry (**Fig. 4**).

Other areas of application include welding of stainless steel to plain carbon steel for outboard motor

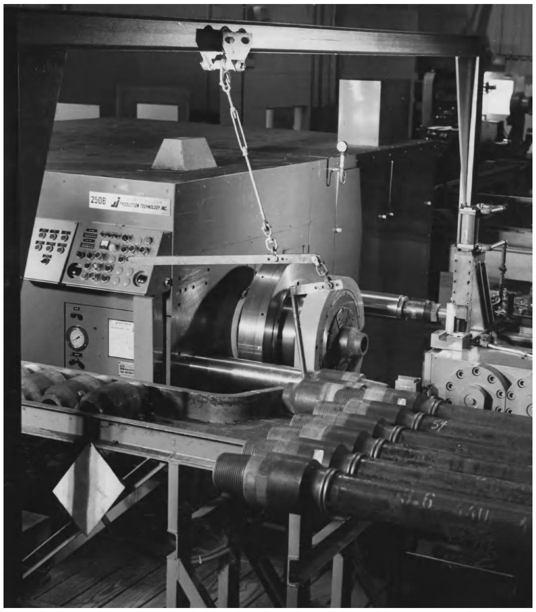


Fig. 4. Inertia welding machine for oil drill pipe sections.

shafts or electric motor shafts. Inertia welding is used to replace electron-beam welding for joining low-alloy steel shafts to nickel-base super-alloy turbine wheels or turbochargers in the aircraft industry. The process is also for making form-cutting tools by welding carbon-steel shanks to tool-steel heads. If one part has a round cross section and the materials are weldable by this process, the number of possible applications is virtually unlimited.

Other developments include design of multiple-spindle machines and radial inertia welding for joining a sleeve onto a ring in the radial direction. *See* METAL, MECHANICAL PROPERTIES OF; WELDED JOINT; WELDING AND CUTTING OF MATERIALS. K. K. Wang

Bibliography. American Welding Society, *Welding Handbook*, 8th ed., vol. 2, 1991; W. M. Hallett, *A Review of Inertia Welding*, SME Pap. AD76-090, 1976; K. K. Wang, *Friction Welding*, Welding Res. Counc. Bull. 204, April 1975.

Inertial navigation system

A self-contained system that can automatically determine the position, velocity, and attitude of a moving vehicle for the purpose of directing its future course. Based on prior knowledge of time, gravitational field, initial position, initial velocity, and initial orientation relative to a known reference frame (coordinate system), an inertial navigation system (INS) is capable of determining its present position, velocity, and orientation without the aid of external information. The generated navigational data are used to determine the future course for a vehicle to follow in order to bring it to its destination. Such systems have found application in the guidance and control of submarines, ships, aircraft, missiles, and spacecraft. The sensors making these measurements, based on one of the basic properties of mass—inertia, are gyroscopes (gyros) and accelerometers. *See* MASS.

Inertial navigation systems are used in many different types of applications that require vastly different performance from the gyros and accelerometers, and they are broadly classified by the quality (accuracy limitations) of their gyros and accelerometers (**Table 1**). Here, the gyro bias drift rate is given in units of degrees/unit time, and the accelerometer bias is given in terms of milli-*g* or micro-*g*, where $1\ g = 9.8\ \text{m/s}^2 = 32\ \text{ft/s}^2$.

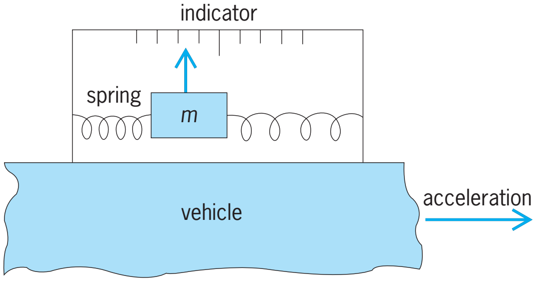


Fig. 1. Elementary accelerometer.

Principle of operation. An inertial navigation system integrates Newton's laws of motion to determine velocity and distance traveled from an initial position in three dimensions (vectorially), and continually updates these calculations. For a vehicle with initial velocity $\mathbf{v_o}$ at distance $\mathbf{s_o}$ from the starting point and acceleration $\mathbf{a}$, the subsequent velocity $\mathbf{v}$ and position $\mathbf{s}$ are given by Eqs. (1) and (2).

$$\mathbf{v} = \mathbf{v_o} + \int \mathbf{a}\,dt \quad (= \mathbf{v_o} + \mathbf{a}\,t \text{ for constant } \mathbf{a}) \tag{1}$$

$$\mathbf{s} = \mathbf{s_o} + \mathbf{v_o}t + \iint \mathbf{a}\,dt^2 \quad (= \mathbf{s_o} + \mathbf{v_o}t + {}^1\!/_2\mathbf{a}\,t^2 \text{ for constant } \mathbf{a}) \tag{2}$$

See ACCELERATION; DYNAMICS; NEWTON'S LAWS OF MOTION; VELOCITY.

The basic measurement of an inertial guidance system is thus acceleration, which is measured by an accelerometer. A notional accelerometer is shown schematically in **Fig. 1**. The acceleration is determined by measuring the inertial force **F** required to

TABLE 1. Classification of inertial navigation systems

Classification	Application	Gyro bias drift rate	Accelerometer bias
Commercial grade	Automobiles	>1°/s	>50 mg
Tactical grade	Short-range missiles	1°/h	1 mg
Navigation grade	Aircraft	0.01°/h	25 μg
Strategic grade	Long-range missiles	0.001°/h	1 μg

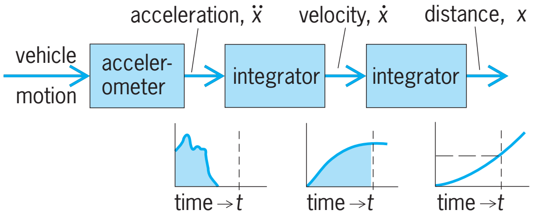


Fig. 2. Computation of distance by two integrations. Integration of acceleration gives velocity; integration of velocity gives distance.

move [or restore to its original (null) position] the proof mass m and inserting this measurement into the expression $\mathbf{a} = \mathbf{F}/m$. However, an accelerometer cannot distinguish between acceleration and gravitation. This can be visualized by considering a vehicle in free fall toward the Earth; an accelerometer would show zero output even though the vehicle is accelerating at 1 g. Conversely, an accelerometer sitting on a fixed table would indicate that the table is accelerating vertically at 1 g, whereas it is actually stationary. This means that gravity (gravitation plus centrifugal acceleration effects from the Earth's rotation) must be compensated in an inertial navigation system. *See* FORCE; RELATIVITY.

If three accelerometers are placed on a stable base and arranged so that their sensitive axes are, respectively, north, east, and down, and the vertical accelerometer is compensated for gravity, then any motion of the base can be sensed and its velocity and the position determined from these three measurements. The velocity is determined by integrating the accelerometer output, and the position is determined by integrating velocity (**Fig. 2**). This idealized concept is complicated by the necessity to maintain the base level and to stay pointed in one direction (leading to a gimbaled inertial system), or alternatively to precisely track the angular motion of the sensor about each axis and computationally determine the effective acceleration in each direction (leading to a strap-down inertial system). In either case, gyroscopes are used to determine the angular motion of the vehicle and to provide the attitude (vector direction) information. Depending on the type of gyro used, the output is either a change in angular rate (rate gyro) or change in angle (integrating gyro). The gyro characteristics, more than any other component, determine the quality of the inertial system. While various types of gyros can be used in either system, generally the floated integrating gyro and the dynamically tuned gyro find their major application in gimbaled systems, and the electrostatic, optical (ring laser and fiber optic), and micromechanical (MEMS) gyros are commonly used in strap-down applications. Depending upon which navigation coordinates are being used, the Earth's rotation (15°/h) has to be compensated in the gyro's output.

A precision clock is required to perform the integration of acceleration to velocity, velocity to distance, and angular rate to angle. *See* ACCELEROMETER; GYROSCOPE.

Gimbaled inertial system. There are several ways that a gimbaled system can be mechanized:

1. The system can be inertially stabilized, in which case the platform is kept fixed relative to inertial space (that is, relative to distant star directions), except for gyro drift.
2. In local level stabilization, the platform is kept aligned with a local level coordinate frame, which rotates with the Earth.
3. In rate biasing, the platform has a large rotation rate about an axis due to deliberate gimbal or vehicle rotation, in order to, for example, prevent laser gyro lock-in.
4. In thrust following, one axis of the platform is kept along the thrust vector of a vehicle.
5. In a carouseled or maytagged system, the platform is continuously rotated about one or more axes to reduce navigation errors due to gyro and accelerometer bias errors.
6. In an indexed system, the platform is rotated in discrete increments about one or more axes to reduce navigation errors by estimating gyro and accelerometer error terms or reducing the effect of bias errors.

For an inertially stabilized three-gimbal platform (**Fig. 3**), the inner gimbal or controlled member is the stable platform on which the three gyros and accelerometers are mounted. As the vehicle to which the system is attached moves, the gyros sense any rotation and drive the gimbals in the opposite direction to null the gyro rate. As a result, the inner platform remains fixed in inertial space. Most inertial platforms are mechanized, however, in the "local vertical" mode, which means that the perpendicular to the stable platform (inner gimbal) always points toward the center of the Earth. Under horizontal accelerations, the pendulum of the accelerometer measuring the vertical direction will tilt and indicate a false vertical. To correct for this, the integrated outputs of the north and east accelerometers (vehicle velocity) are divided by the radius of the Earth and used as an additional signal to drive the gimbal torquers.

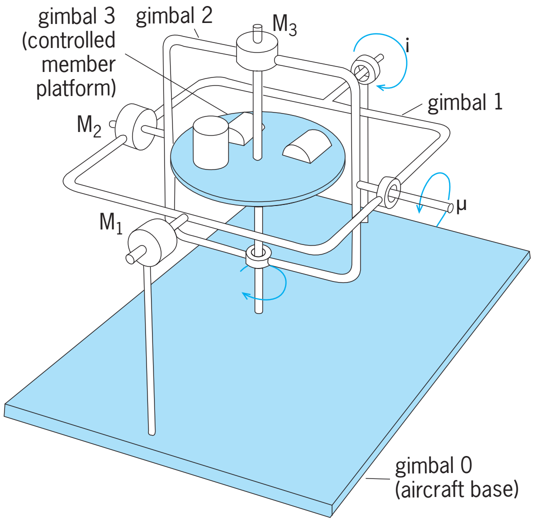


Fig. 3. Schematic of a three-gimbal platform.

This is known as a Schuler loop. If the output errors of such a system are observed, they will contain an oscillation with a period of the order of 84 min, known as the Schuler frequency. The vertical position channel of a stable platform is unstable; hence, the vertical accelerometer is sometimes deleted. *See* SCHULER PENDULUM.

Strap-down inertial system. In a strap-down inertial system, the three accelerometers and three gyros are hard-mounted to the vehicle (**Fig. 4**). As the vehicle moves, the gyro and accelerometer outputs are read approximately every 1/100 of a second, the new direction is established, and the velocity change and position change are calculated for the past 1/100 s. These changes are then summed with previous data to give total velocity and position changes. The computational burden is very heavy, especially for the coordinate transformations necessary to maintain the mathematical equivalent of the gimbal position in inertial space. Strap-down systems also need to be Schuler-tuned, where this operation is accomplished mathematically in the navigation computer. Two major contributors to the practicality of the modern strap-down system were the solid-state high-speed computer and the development of gyros that do not require torquers and have outputs that are insensitive to vehicle accelerations, such as the ring laser gyro. Nearly all applications now use strap-down systems. However, at present the most precise navigators are still gimbaled systems. In a strap-down system, the "vertical accelerometers" cannot be deleted since the accelerometers are tied to the vehicle and may assume any attitude. A schematic of the inertial navigation process for a strap-down system navigating in the local level (rotating with the Earth) coordinate frame is shown in **Fig. 5**.

Typical aircraft-quality strap-down navigators are relatively large (around 7500 cm^3 or 450 in.^{3}) and expensive. Navigators for short-range missiles and artillery shells are much smaller (150–500 cm^3 or 10–30 in.^{3}). **Figure 6** shows examples of strap-down inertial navigation systems.

Other terminologies. Navigation is only part of the process of arriving at the desired destination. The first requirement for achieving this objective is a knowledge of where you are and where your destination is (using maps or surveys). The second is navigation (using sensors or a navigation system). The third is guidance—that is, based on the navigation information, what needs to be done to get to the destination or target (using a guidance computer). Finally, there is the control stage, in which the guidance commands are implemented (using actuators). Often guidance and navigation are used interchangeably. Other terminologies commonly used in navigation are given in **Table 2**. *See* GUIDANCE SYSTEMS.

Aiding. Most inertial navigation systems incorporate some form of aiding (also called augmentation) to improve their accuracy by periodically updating the inertial navigation system to overcome its drift over time. Typical aiding systems could be one or more of the following: airspeed indicator, magnetic compass, magnetometer, odometer, baroaltimeter, infrared (IR) seeker, radar, Doppler radar, star tracker, Loran, or Global Positioning System (GPS). Kálmán filtering techniques are used to update the navigation state to provide the "optimal" solution. **Figure 7** shows a schematic of an automobile navigation system with several augmentation systems. *See* AIRSPEED INDICATOR; ALTIMETER; DOPPLER RADAR; ESTIMATION THEORY; INFRARED RADIATION; LORAN; MAGNETIC COMPASS; MAGNETOMETER; RADAR; SATELLITE NAVIGATION SYSTEMS; STAR TRACKER.

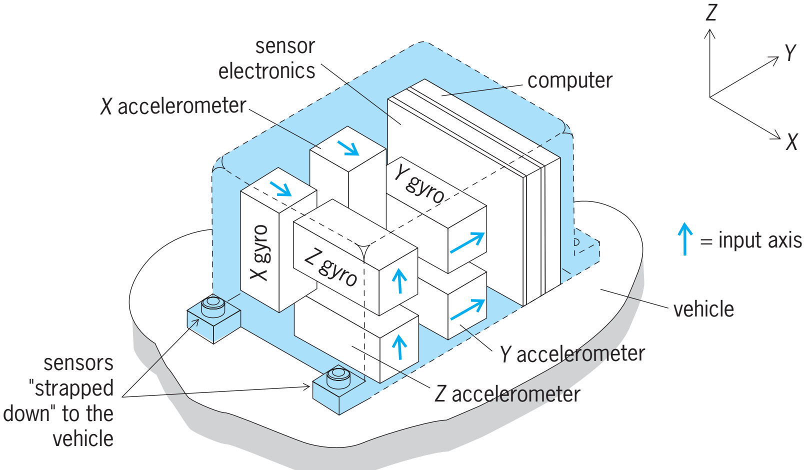


Fig. 4. Strap-down inertial navigation system. (*After A. Lawrence, Modern Inertial Technology, 2d ed., Springer, 1998*)

System initialization. When an inertial system is first turned on, it must be initialized. This is commonly done by electronically inserting the present latitude and longitude into the navigation computer via a key pad or from the GPS signal. The inertial unit itself must determine the vertical since its precise alignment relative to the vehicle is unknown, as is the precise alignment of the vehicle itself. In a gimbaled system, the vertical is determined by nulling the north and east accelerometer signals (that is, leveling the platform). The north alignment is done by

TABLE 2. Navigation terminologies

Term	Definition
Inertial sensor assembly (ISA)	A structure containing multiple inertial sensors in fixed orientations relative to one another
Attitude heading reference system (AHRS)	Estimates the body angles of a vehicle with respect to the local level coordinate system; performs direction finding (for example, north), but is not a navigator
Inertial reference unit (IRU)	Measures angular motion in three dimensions without external reference (usually used in satellites)
Inertial measurement unit (IMU)	Measures linear and angular motion in three dimensions without external reference (provides output from an ISA to the navigation computer)

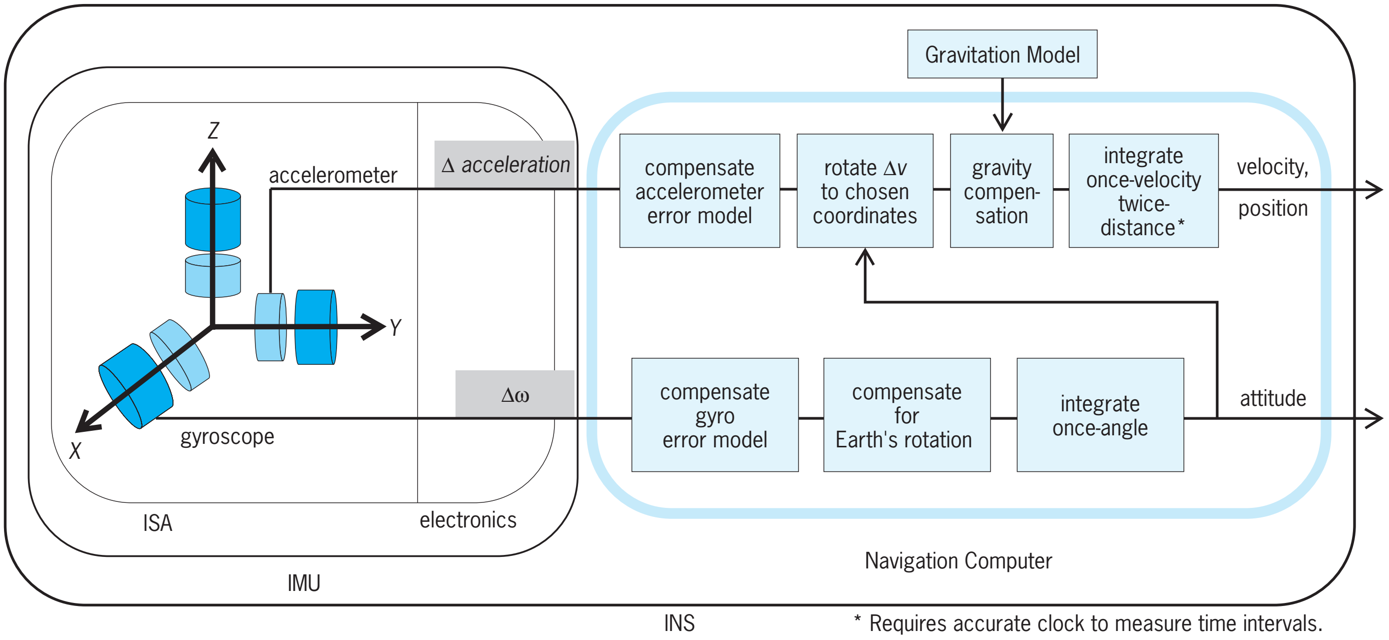


Fig. 5. Inertial navigation process. The inertial sensor assembly (ISA) and inertial measurement unit (IMU) are defined in Table 2.

gyrocompassing the platform until no Earth rate is sensed by the east gyro and also until the north and east accelerometers have zero output. This process may take from 10 to 30 min, depending upon the quality of the instruments and the vehicle's latitude. Because of this alignment technique, the inertial system aligns north instead of magnetic north in order to be parallel to the Earth's spin vector. In strap-down systems, the mathematical equivalent of aligning the gyros and accelerometers takes place in the computer. Down is defined to be along the accelerometer triad vector (gravitation), east is defined to be along the vector cross product of the gyro triad vector (Earth's rotation) and up, and north is defined to be the vector cross product of the first two. In many military systems, the vehicle is initialized with a transfer alignment from a more accurate navigation system (for example, from an aircraft's navigation system to the inertial navigation system of the missile being carried). *See* CALCULUS OF VECTORS.

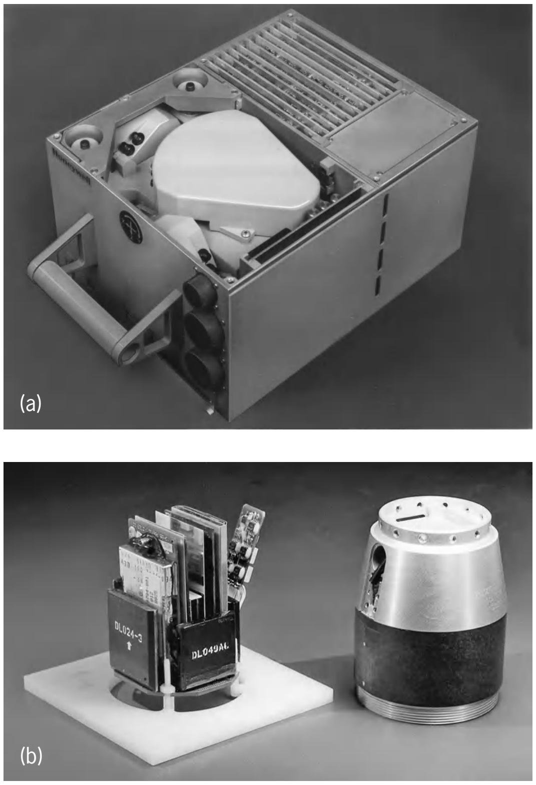


Fig. 6. Strap-down inertial navigation systems. (*a*) Mock-up of system based on laser gyro technology. The three laser gyros, hermetically sealed, are on the left (*Honeywell*). (*b*) 130-cm^3 (8-in.^{3}) system that uses silicon MEMS technology (*Draper Laboratory*).

System accuracies. The accuracies of aircraft inertial systems are usually stated as a drift rate in nautical miles per hour. Typical commercial systems such as those that fly in transoceanic airliners have accuracies slightly better than 1 nautical mile per hour (1.85 km/h); therefore, in a 5-h flight the total expected error would be less than 5 nmi (9.3 km), which would allow other navigation aids to be acquired at the destination. The major contributor to this error is gyro drift; as a rule of thumb, a gyro with a drift rate of 0.01° per hour could be expected to produce an error of 1 nmi/h when incorporated into an inertial system. Below a capability of approximately 0.25 nmi/h (0.5 km/h), the uncertainties in the gravitational field vector begin to be significant sources of error. Nevertheless, systems have been built which consistently exhibit errors less than this.

For a short-range-missile inertial navigation system with an accelerometer error of 1 milli-*g* and perfect gyros, the distance error over a 1-min flight would be around 20 m (60 ft).

Aiding the navigation system by using information from the GPS allows improved navigation accuracy when using lower-performing gyros and

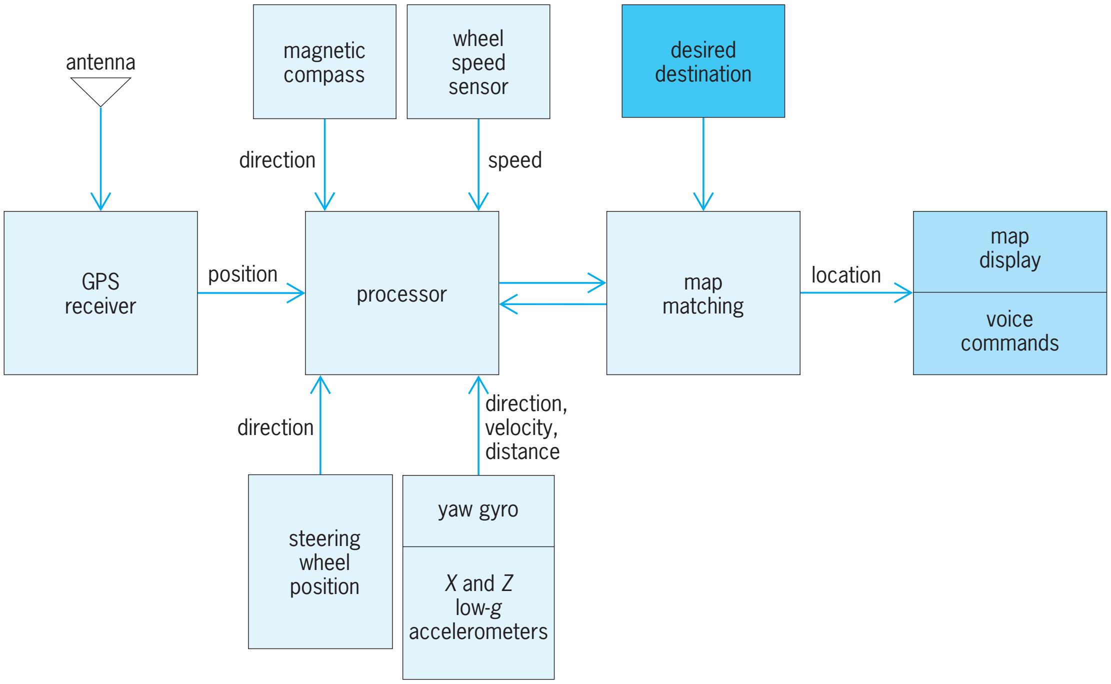


Fig. 7. Automobile navigation system.

accelerometers, and has been responsible for the introduction of inertial navigation to many new applications, such as automobiles and guided bombs. Each technology counteracts the weakness of the other. GPS errors are bounded, but updates are available at a lower rate, and there is the possibility of jamming or of signals becoming unavailable (for example, in parts of city centers or in buildings). Inertial navigation system errors are unbounded (grow over time), but they provide higher rate updates, and cannot be jammed. Also, the inertial navigation system can perform the autopilot function (stabilization and control) for the vehicle. Most navigation systems are no longer purely inertial but are now INS/GPS systems, and would not be able to achieve the desired performance with the inertial navigation system alone. The 1-nmi/h aircraft navigator described above, when coupled with GPS, would be accurate to the order of meters per hour. Neil M. Barbour; William E. Howell

Infant diarrhea

Diarrhea and its complications are the most important causes of infant death in most developing regions. The causes of the illness vary from dietary incompatibilities to intestinal infection. The most important infectious causes are, in approximate order of importance, rotavirus, the bacteria *Campylobacter jejuni*, *Shigella* (often causing dysentery or passage of bloody stools), *Salmonella*, and *Escherichia coli*, and the parasites *Giardia lamblia*, *Cryptosporidium*, and *Entamoeba histolytica*. The microorganisms are spread to the infant from other persons or through food or water. Breast-feeding is associated with a decreased occurrence of diarrhea and represents a major means of preventing infantile diarrhea in the developing world.

Since dehydration is the complication which must be prevented in diarrhea, attention to fluid and electrolyte replacement is the cornerstone of therapy. Oral electrolyte solutions administered under supervision are generally effective even in severely dehydrating forms of diarrhea. In a small percentage of cases, oral fluids cannot keep up with body losses, particularly if vomiting complicates the illness. In these situations, hospitalization and intravenous fluid and electrolyte replacement are lifesaving. Antibiotics are useful in the therapy of infants with more serious forms of intestinal infection. Hospital nursery outbreaks represent serious public health problems, occurring because of the appearance of infectious organisms in the hospital environment together with the presence of a confined nursery population of infants which are highly susceptible to these agents. *See* BACILLARY DYSENTERY; DIARRHEA; GIARDIASIS. Herbert L. Dupont

Infant respiratory distress syndrome

A common disorder of premature (or preterm) birth, defined by respiratory difficulty, which requires oxygen or assisted ventilation, and characteristic changes on chest x-ray. It is primarily the result of deficiency in surface-active lipids (contained within pulmonary surfactant) which serve to stabilize the air spaces (alveoli) within the lung. Although most cases of infant respiratory distress syndrome rarely result in a fatal outcome in developed countries, all require specialized care in a neonatal intensive care unit.

Incidence and pathophysiology. The incidence of respiratory distress syndrome increases with greater prematurity as well as with maternal diabetes and asphyxia around the time of delivery. In contrast, growth retardation, other prolonged stress in the uterus, and maternal steroid administration decrease the incidence of infant respiratory distress syndrome. In addition, male infants are more predisposed to the complications of prematurity, including infant respiratory distress syndrome. *See* DIABETES.

The disease was initially termed hyaline membrane disease because autopsies demonstrated a membranous material lining the air spaces of these lungs, which are also very solid secondary to poor aeration. However, these membranes are not the cause of poor gas exchange, but are the result of lung injury which produces a layer of debris from blood products and injured lung epithelial cells. Rather, the primary cause of respiratory distress syndrome is deficiency of pulmonary surfactant. Pulmonary surfactant is manufactured in specialized lung cells and secreted to line the surfaces of alveoli. It has unique physical properties that lower surface tension and prevent the alveoli from collapsing. Surfactant deficiency and the resultant collapse of alveoli cause the lung to fail in its primary role of oxygen uptake and carbon dioxide elimination. Severe untreated infant respiratory distress syndrome therefore results in respiratory failure. Moreover, failure of lung aeration has cardiovascular consequences secondary to constriction of pulmonary blood vessels and decrease in blood flow to the lungs. Respiratory failure results in poor oxygenation of tissues and impaired metabolism throughout the body.

Clinical features. Infants with respiratory distress syndrome show signs of respiratory difficulty (including rapid respiration, retractions of the rib cage, nasal flaring, grunting or whining upon exhaling, and a bluish discoloration of the skin and mucous membranes) if adequate supplemental oxygen is not provided. Confirmation of the diagnosis using radiographic methods is essential, because a variety of other disorders (such as infection) may have a similar clinical picture. A characteristic x-ray shows a uniform, diffuse haziness of both lungs due to their poor aeration. Biochemical confirmation of surfactant deficiency in lung secretions would ideally confirm the diagnosis; however, this is not practical or widely available. Infant respiratory distress syndrome is usually well advanced 5–7 days after birth; however, complications of the disease and its therapy may prolong this course in the smallest and sickest babies.

Prevention and treatment. Prevention of prematurity and administration of good prenatal care reduce the incidence of infant respiratory distress syndrome. If fetal well-being is in jeopardy and the mother must undergo induced labor before the end of her term, pulmonary maturity of the fetus can be determined by analysis of amniotic fluid (which is in part composed of fetal lung fluid), sampled by means of amniocentesis. Corticosteroids given before birth to the mother at risk for premature delivery is a well-established therapy for reducing the incidence of respiratory distress syndrome; this treatment enhances pulmonary maturation in the infant.

Surfactant therapy is now widely used for treatment of infants with moderate or severe respiratory distress syndrome. This requires introducing a tube into the trachea for direct placement of surfactant into the infant's lungs; this is done one or more times over the first 48 h of life. Both synthetic and animal based preparations of surfactant are effective in reducing the mortality of preterm infants, and are free of known complications.

Surfactant therapy must be performed in conjunction with conventional treatment such as maintenance of an optimal-temperature environment and use of intravenous fluids. Oxygen is administered directly to the spontaneously breathing infant or by means of assisted ventilation. Blood gas status is closely checked by noninvasive monitoring combined with frequent blood sampling. Throughout the course of respiratory distress syndrome, complications of this disease, its management, or prematurity in general may arise at any time.

Prognosis. Infants with respiratory distress syndrome can be successfully treated, especially with the widespread use of surfactant therapy. However, the outcome remains heavily influenced by the infant's degree of prematurity. Very immature infants and those with severe respiratory distress syndrome may develop a more prolonged form of neonatal respiratory disease called bronchopulmonary dysplasia. This has been attributed to lung injury caused by exposure of an immature respiratory system to high inspired oxygen. Thus, the most severe cases of infant respiratory distress syndrome, requiring the most intensive oxygen and ventilator therapy, are consequently the most prone to prolonged lung injury. Fortunately only a small proportion of these infants remain symptomatic throughout childhood. *See* RESPIRATORY SYSTEM DISORDERS.

Richard J. Martin

Bibliography. M. E. Avery and J. Mead, Surface properties in relation to atelectasis and hyaline membrane disease, *Amer. J. Dis. Child.*, 97:517–523, 1959; A. A. Fanaroff and R. J. Martin (eds.), *Neonatal-Perinatal Medicine*, 5th ed., 1992; T. Fujiwara et al., Artificial surfactant therapy in hyaline membrane disease, *Lancet*, 1:55–59, 1980; A. H. Jobe, Pulmonary surfactant therapy, *N. Engl. J. Med.*, 328:861–868, 1993; M. H. Klaus and A. A. Fanaroff (eds.), *Care of the High-Risk Neonate*, 5th ed., 2001.

Infarction

The process of anoxic tissue death. The usual cause is occlusion of an artery by a thrombus or embolus and sometimes by severe atherosclerosis. The development of the infarct depends to a great extent on the collateral circulation. If the collateral blood supply is inadequate or if the vessel is the sole source of blood supply to the region, an infarct results. *See* ARTERIOSCLEROSIS; THROMBOSIS.

Following an initial period of hyperemia, the infarcted region becomes pale. Because of the branching arrangement of the vascular system, infarcts of the spleen, kidney, and lung are usually wedge-shaped, with the base of the wedge at the periphery and the apex toward the point of the vascular obstruction (**Fig. 1**).

With occlusion of an artery the existing collateral

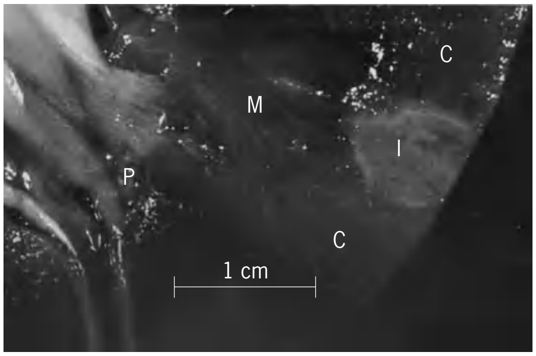


Fig. 1. Magnified gross photograph of a wedge-shaped infarct of kidney (I). Renal cortex (C), medullary pyramid (M), and renal pelvis (P) can be seen.

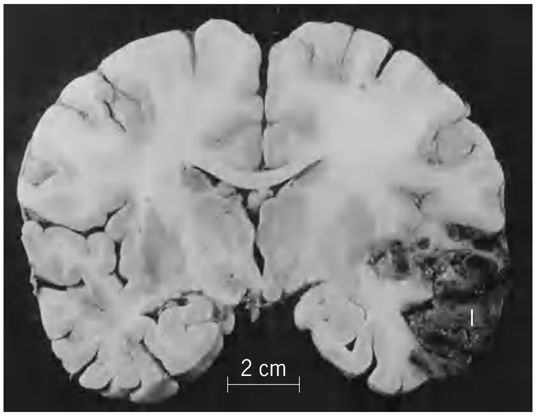


Fig. 2. Gross photograph of a brain section, showing a wedge-shaped infarct (I) that resulted from a thrombus. Secondary hemorrhage has occurred.

circulation attempts to bring more blood to the region. Depending on the extent of the process, the infarcted area will be red or pale. This process varies with the organ involved. Those of the kidney and heart are pale, whereas those of the spleen and lung are usually red. With death of the tissue a secondary hemorrhage may occur (**Fig. 2**).

The tissue in the affected area undergoes coagulative necrosis and dies while the tissue around the infarcted region becomes inflamed. In time the dead tissue is replaced by a shrunken depressed scar. *See* INFLAMMATION.

Infarcts commonly occur in the lungs, heart, brain, spleen, and kidneys. The common cause of infarcts of the heart is thrombosis of the coronary artery, usually secondary to atherosclerosis. Hemorrhage into an atherosclerotic plaque can also result in thrombosis of an artery.

Myocardial infarction is usually accompanied by severe prolonged substernal oppression or pain, electrocardiographic changes, fever, leukocytosis, and serum enzyme changes. Myocardial infarcts usually involve the left ventricle or interventricular septum and only rarely involve the right ventricle or atria. The location of the infarct usually depends on the coronary artery occluded. If death does not occur immediately, the infarct heals and is replaced by scar tissue in 5–8 weeks, depending on its size.

Complications of myocardial infarction include rupture of the necrotic portion of the heart with hemorrhage into the pericardium and rapid death; mural thrombi which form on the endocardial surface of the healing infarct and which may give rise to embolism; congestive heart failure as a result of an insufficient amount of remaining healthy muscle to carry on the work of the heart (occurring most commonly with repeated infarcts); and disturbances in heart rhythm if the pacemaker or conduction system is involved.

Infarction of a portion of bowel will result in death unless surgical intervention is forthcoming. Embolization to the lungs is a rather frequent occurrence. However, because of the collateral blood supply of the lungs, infarction follows only when there is some interference with the circulation, such as chronic pulmonary venous congestion. An extensive collateral circulation also exists in the liver, hence the rarity of infarcts in this organ. *See* CIRCULATION DISORDERS.

Romeo A. Vidone; Irwin Nash

Bibliography. E. Braunwald et al. (eds.), *Harrison's Principles of Internal Medicine,* 15th ed., 2001; R. S. Cotran et al., *Robbins' Pathologic Basis of Disease*, 6th ed., 1999; J. M. Kissane (ed.), *Anderson's Pathology*, 9th ed., 1989; M. D. Silver, *Cardiovascular Pathology*, 2 vols., 3d ed., 2000.

Infection

A term considered by some to mean the entrance, growth, and multiplication of a microorganism (pathogen) in the body of a host, resulting in the establishment of a disease process. Others define infection as the presence of a microorganism in host tissues whether or not it evolves into detectable pathologic effects. The host may be a bacterium, plant, animal, or human being, and the infecting agent may be viral, rickettsial, bacterial, fungal, or protozoan.

A differentiation is made between infection and infestation. Infestation is the invasion of a host by higher organisms such as parasitic worms. *See* EPIDEMIOLOGY; HOSPITAL INFECTIONS; MEDICAL BACTERIOLOGY; MEDICAL MYCOLOGY; MEDICAL PARASITOLOGY; OPPORTUNISTIC INFECTIONS; PATHOGEN; VIRUS.

Daniel N. Lapedes

Infectious disease

Infectious diseases are any of those pathological conditions spread among all biological species. Infectious diseases, although varied in their effects, are always associated with viruses, bacteria, fungi, protozoa, multicellular parasites and aberrant proteins known as prions. A complex series of steps, mediated by factors contributed by both the infectious agent and the host, is required for microorganisms or prions to establish an infection or disease. Worldwide, infectious diseases are the third leading cause of human death.

Pathogenic microorganisms. It is difficult to group microorganisms as exclusively commensal or pathogenic species. The most common relationship between a host and a microorganism is a commensal one, in which advantages exist for both organisms. For example, hundreds of billions of bacteria of many genera live in the human gastrointestinal tract, coexisting in ecological balance without causing disease. These bacteria help prevent the invasion of the host by more virulent organisms such as *Salmonella*, *Shigella*, or *Campylobacter* species. In exchange, the host provides an environment in which harmless bacteria can readily receive nutrients to fulfill their metabolic needs. There are very few microorganisms that cause disease every time they encounter a host. Instead, many factors of both host and microbial origin are involved in infectious disease. These factors include the general health of the host, previous exposure of the host to the microorganism, and the complement of molecules produced by the bacteria. Pathogenic microorganisms are frequently categorized by the host site where they most commonly cause disease. For example, *Bacteroides*, *Prevotella*, and *Haemophilus* species are usually associated with respiratory tract infections, while poxviruses and *Trichophyton* species are frequently encountered on the skin. Some microorganisms, such as *Staphylococcus* species, can cause a myriad of infections involving the respiratory, gastrointestinal, and genitourinary tracts, as well as skin, eyes, or ears.

Infectious disease transmission. Spread of a pathogenic microorganism among individual hosts is the hallmark of an infectious disease. This process, known as transmission, may occur through four major pathways: contact with the microorganism, airborne inhalation, through a common vehicle such as blood, or by vector-borne spread.

Contact spread. In contact-spread disease, the host is in direct or indirect contact with the infectious agent. Most sexually transmitted diseases are spread by direct contact with the microorganism from an infected host. Indirect transmission occurs via an inanimate object that carries the infecting organism, as in spread of infections resulting from the sharing of needles among intravenous drug abusers. Another route of transmission involves spread by droplets, relatively large particles coughed or sneezed into the air. The droplets are usually broadcast within a few feet of the infected host and transmitted to a susceptible individual nearby; examples include streptococcal pharyngitis (strep throat) and measles. *See* STREPTOCOCCUS.

Airborne spread. Transmission through the air (beyond a few feet from the infected host) occurs when microorganisms attach to dust particles or when they are contained in aerosols. Particles may remain suspended in the air for hours or days, and if they are of the appropriate size they can be inhaled. Tuberculosis (*Mycobacterium tuberculosis*) and chickenpox (varicella zoster virus) are examples of diseases transmitted by the airborne route. *See* CHICKENPOX AND SHINGLES; TUBERCULOSIS.

Common-vehicle spread. Persons can become infected following contact with a contaminated common vehicle, such as food, water, or blood. Poisoning by *Escherichia coli* toxins, mycotoxins (fungal toxins), *Clostridium botulinum* toxins, and the *Staphylococcus aureus* enterotoxins is primarily through the ingestion of contaminated food products. The infectious agents of spongiform encephalopathies, such as mad cow disease, are known as prions and are transmitted through contaminated feed. Typhoid fever, cholera, and amebic dysentery are examples of illnesses acquired through contaminated drinking water. Hepatitis B and acquired immune deficiency syndrome (AIDS) are examples of viral diseases spread by transfusion with infected blood. *See* BOTULISM; FOOD POISONING; HEPATITIS; WATERBORNE DISEASE.

Vector-borne spread. Transmission of microorganisms can occur through an intermediate vector, such as an insect of arthropod. Often the intermediate host is a necessary agent, serving an essential biological role in the microorganism's life cycle. For example, mosquitoes spread the malaria-causing protozoon *Plasmodium falciparum* and the yellow fever arbovirus as they transmit infected blood from one mammal to another. Spotted fever, caused by *Rickettsia rickettsii*, and Lyme disease, caused by the spirochete *Borrelia burgdorferi*, are transmitted via the bite of an infected tick. *See* LYME DISEASE; MALARIA; YELLOW FEVER.

Pathogenesis of infection. The manner in which an infectious disease develops, or its pathogenesis, usually follows a consistent pattern. To initiate an infection, there must be a physical encounter as which the microorganism enters the host. The most frequent portals of entry are the respiratory, gastrointestinal, and genitourinary tracts as well as breaks in the skin. Surface components on the invading organism determine its ability to adhere and establish a primary site of infection. For example, the influenza A virus targets epithelial cells of the respiratory tract. After infection, the virus interacts with the host to cause typical symptoms of the flu. In another example, the protozoon *Entamoeba histolytica* adheres to cells lining the colon to produce the diarrhea associated with amebic dysentery. The cellular specificity of adherence of microorganisms often limits the range of susceptible hosts. For example, although measles and distemper viruses are closely related, dogs do not get measles and humans do not get distemper.

From the initial site of infection, microorganisms may directly invade further into tissues or travel through the blood or lymphatic system to other organs. For example, *Neisseria meningitidis* colonizes the mucous membranes of the upper respiratory tract of humans. However, the bacteria must spread through the bloodstream into the protective membranes that cover the brain and spinal cord for a life-threatening disease known as meningitis to occur.

Microorganisms produce toxins that can cause tissue destruction at the site of infection, can damage cells throughout the host, or can interfere with the normal metabolism of the host. The damage that microorganisms cause is directly related to the toxins they produce. Toxins are varied in their mechanism of action and host targets. For example, *Staphylococcus aureus* can express several toxins that destroy tissue near its site of entry into the body, leading to the formation of boils, carbuncles, and abscesses. The tissue-damaging toxins collegenase, coagulase, hemolysins, proteases, and hyaluronidases serve either to protect the bacteriaL host defenses or to promote the spread of bacteria within the host. For example, coagulase, in conjunction with serum factors from the host, contributes to the formation of a fibrin clot that surrounds the bacteria, insulating them from destruction by the host. Alternatively, the enzyme hyaluronidase can weaken the host's connective tissue, facilitating the spread of the bacteria. Another bacterium, *Clostridium tetani*, produces a powerful toxin with a systemic effect. This toxin blocks the normal relaxation impulses of the nerves and thus causes the severe muscle spasms characteristic of tetanus. Cholera can interfere with the metabolism of the host. This ancient disease is contracted by ingesting water contaminated by the bacterium *Vibrio cholerae*. The bacterium releases a protein (an enterotoxin) that causes severe diarrhea by stimulating the secretion of enormous quantities of watery fluid from the small intestine. Patients lose so much body fluid that they go into shock and die. *See* CHOLERA; STAPHYLOCOCCUS; TETANUS.

Response to infection. The host's reaction to an infecting organism is the inflammatory response, the body's most important internal defense mechanism. Although the inflammatory response is also seen as secondary to physical injury and nonspecific immune reactions, it is a reliable indicator of the presence of pathogenic microorganisms. Inflammation is characterized by the cardinal signs of heat, redness, swelling, and pain. The warmth and redness are due to the dilation of minute blood vessels near the site of infection. The increased permeability of those vessels causes them to release fluid into the area causing swelling, and the irritation of nearby nerve endings causes pain. The dilated small vessels allow more blood to flow to the site of infection, while the enhanced permeability of the blood vessels permits increased numbers of disease-fighting substances to the area. Immune cells known as lymphocytes and granulocytes are carried by the blood to the site of infection. These cells either engulf and kill, or secrete substances which inhibit and neutralize, microorganisms. Other white blood cells, primarily monocytes, recognize foreign organisms and transmit chemical signals to other cells of the host's immune system, triggering the production of specific antibodies or specialized killer cells, both of which are lethal to the infecting microorganism. Any influence that reduces the immune system's ability to respond to foreign invasion, such as radiation therapy, chemotherapy, or destruction of immune cells by an immunodeficiency virus such as HIV, increases the likelihood that a organism will cause disease within the host. *See* INFLAMMATION.

Infectious disease control. Chemical compounds that are more toxic to microorganisms than to the host are commonly employed in the prevention and treatment of infectious disease; however, the emergence of drug-resistant organisms has led to increases in the morbidity and mortality associated with some infections. Other methods for controlling the spread of infectious diseases are accomplished by breaking a link in the chain of transmission between the host, microorganism, and mode of spread by altering the defensive capability of the host. Overall, the three most important advances to extend human life are clean water, vaccination, and antibiotics (in that order of importance).

Vector control. Water-borne infections are controlled by filtration and chlorination of municipal water supplies. Checking food handlers for disease, refrigeration, proper cooking, and eliminating rodent and insect infestation have markedly reduced the level of food poisonings. The transmission of vector-borne diseases can be controlled by eradication of the vector. Control of malaria in the United States succeeded after mosquito-breeding areas were drained and sprayed with insecticides. Blood-borne infections are reduced by screening donated blood for antibodies specific for HIV and other viruses and by rejecting donations from high-risk donors. For diseases such as tuberculosis, the airborne spread of the causative agent, *Mycobacterium tuberculosis*, can be reduced by quarantining infected individuals. The spread of sexually transmitted diseases, including AIDS, syphilis, and herpes simplex, can be prevented by inhibiting direct contact between the pathogenic microorganism and uninfected hosts. *See* ACQUIRED IMMUNE DEFICIENCY SYNDROME (AIDS).

Vaccination. The spread of an illness can be halted by activating the immune system of the host. The most beneficial implementation of that concept has been vaccination, which renders an individual immune to an illness. Children in the United States are routinely immunized against diphtheria, whooping cough, tetanus, measles, mumps, rubella, chickenpox, and poliomyelitis. Because of a global vaccination programs, smallpox has been eradicated, and soon polio may be eliminated throughout the world.

Antimicrobial drugs. Drug therapy encompasses many classes of compounds that are active against common pathogens. These drugs are the "magic bullets" envisioned by Paul Erhlich to kill bacterial, viral, fungal, or protozoan pathogens without harming their human host. The pencillins, cephalosporins, and vancomycin are selective inhibitors of synthesis of the rigid cell wall that surrounds bacteria. Other antibacterial drugs, such as the tetracyclines and the aminoglycosides, block protein synthesis by the bacteria but not by the cells of the host. Still other compounds can alter the cell wall of bacterial cells, blocking the transport of nutrients into the cell. Viruses, because of their unique structure and method of replication, can be susceptible to drugs such as amantadine hydrochloride, which blocks the penetration of certain myxoviruses such as influenza A, or drugs that inhibit enzymes such as deoxyribonucleic acid (DNA) polymerases, proteases, or kinases that viruses use to replicate. Compounds that inhibit or suppress the growth of pathogenic fungi and protozoa have been used as effective treatment for infectious diseases caused by these agents, although side effects can limit their utility. *See* DIPHTHERIA; EPIDEMIOLOGY; MUMPS; POLIOMYELITIS; RUBELLA; SMALLPOX; VACCINATION; WHOOPING COUGH. Peter J. McNamara

Bibliography. J. V. Bennet and P. S. Brachman, *Hospital Infections*, 2d ed., 1986; P. D. Hoeprich, *Infectious Diseases: A Modern Treatise of Infectious Processes*, 3d ed., 1983; P. R. Murray et al. (eds.), *Manual of Clinical Microbiology*, 6th ed., 1995; G. P. Youmans, P. Y. Paterson, and H. M. Sommers, *The Biological and Clinical Basis of Infectious Diseases*, 3d ed., 1986.

Infectious mononucleosis

A disease of children and young adults, characterized by fever and enlarged lymph nodes and spleen. Only in the mid-1970s did accumulated evidence permit clear linkage of this disease to EB (Epstein-Barr) herpesvirus as the causative agent.

Onset of the disease is slow and nonspecific with variable fever and malaise; later, cervical lymph nodes enlarge, and in about 50% of cases the spleen also becomes enlarged. The disease lasts 4–20 days or longer. Epidemics are common in institutions where young people live. EB virus infections occurring in early childhood are usually asymptomatic. In later childhood and adolescence, the disease more often accompanies infection—although even at these ages inapparent infections are common.

Seroepidemiologic studies indicate that infection with EB virus is common in different parts of the world. In some areas, including the urban United States, about 50% of children 1 year old, 80–90% of children over age 4, and 90% of adults have antibody to EB virus. The mechanism of virus transmission is unknown, but it could take place through oropharyngeal excretion, which has been found to be very common, especially in those with overt disease, and can occur in the absence of symptoms and be prolonged.

For diagnosis, the heterophile antibody test is useful. This test is based on a nonspecific serologic reaction present at high levels in patients with infectious mononucleosis. Total white blood cell count and differential blood count are also useful in diagnosis. *See* CLINICAL PATHOLOGY; HETEROPHILE ANTIGEN.

The nature of the immune response, which terminates the disease and the shedding of virus, is not fully understood; the virus appears to persist in the host at a location inaccessible to neutralizing antibody. Unlike short-lived heterophile antibodies, neutralizing antibodies against EB virus persist for years.

The presence of EB virus, an antigenically distinct member of the herpes group, was initially detected by means of electron microscopy in a small proportion of cells in continuous lymphoblastoid cell lines derived from Burkitt's lymphoma, a tumor indigenous to children in central Africa. EB virus has been detected in peripheral blood leukocytes of persons with infectious mononucleosis, leukocytes from normal individuals, and also in lymphoblastoid cell lines derived from nasopharyngeal carcinomas. A causative role of the virus in the cancers has not been unequivocally shown; it is possible that the virus is merely a "passenger" that grows well in the type of cell in the neoplasm. *See* EPSTEIN-BARR VIRUS; HERPES. Joseph L. Melnick

Bibliography. A. S. Evans (ed.), *Viral Infections of Humans*: *Epidemiology and Control*, 4th ed., 1997; G. Miller, Epstein-Barr herpesvirus and infectious mononucleosis, *Prog. Med. Virol.*, 20:84–112, 1975.

Infectious myxomatosis

A viral disease of European rabbits (*Oryctolagus*) and domestic rabbits, spread mainly by biting insects (mosquitoes and rabbit fleas) and characterized by edematous swellings of the skin, particularly on the head and anogenital area. The disease is caused by infection with myxoma virus, a pox virus, which occurs naturally in certain species of the genus *Sylvilagus* in North, Central, and South America. In these rabbits, infection results generally in localized, nonmalignant tumors that disappear in weeks or months.

Myxomatosis was introduced into Australia in 1950 as a method of controlling rabbits, which were causing great economic and environmental damage. It was also introduced into France in 1952 and spread throughout western Europe, including the British Isles. Although the virus strains introduced into Australia and France killed virtually every rabbit infected and rabbit populations were dramatically reduced, they were not eradicated, as less virulent strains appeared. Myxomatosis now has much less effect on rabbit populations: one estimate is that only 12–19% of rabbits die from the disease.

In typical wild rabbit populations, most of the adults are immune, following recovery from earlier infection. In domestic and commercial rabbits, there is no buildup of immunity or resistance, because of the common policy of isolating and slaughtering infected animals. There is no effective treatment once clinical signs have appeared. Preventive measures include restriction of contact with insect vectors and vaccination with active or inactivated myxoma virus or with Shope fibroma virus. *See* ANIMAL VIRUS. John Ross

Bibliography. F. Fenner and F. N. Ratcliffe, *Myxomatosis*, 1965; H. V. Thompson and C. M. King, *The European Rabbit*, 1994.

Infectious papillomatosis

A nonfatal viral disease that occurs naturally in cottontail rabbits in states bordering the Mississippi River as well as in Oklahoma and Texas, and in brush rabbits in California. Cottontail rabbits from other parts of the United States and domestic rabbits (*Oryctolagus*) are also susceptible. The disease is caused by infection with Shope papilloma virus, one of a number of papilloma viruses that infect a range of animals, including humans; it is spread naturally by contamination of broken skin or by the rabbit tick.

The disease is characterized by cutaneous warts (papillomas) which can persist for months or years, sometimes becoming very prominent. Some papillomas eventually disappear, but in 25% of infected cottontail rabbits and 75% of infected domestic rabbits these warts become malignant carcinomas.

Rabbits can be immunized against papillomatosis by injection of active or inactivated virus. The virus can readily be recovered from papillomas on cottontail rabbits but rarely from those on domestic rabbits, indicating that it exists in mature fully infective virus and immature noninfective virus, which will proliferate in the host and stimulate an immune response. The disease is unlikely to spread between domestic rabbits because infection does not produce infective virus. *See* ANIMAL VIRUS. John Ross

Infertility

Inability to conceive or induce conception. Of all cases of infertility, 35% may be attributed to the male and 55% to the female; the remaining 10% is undetermined.

The female. The principal cause of increasing rates of infertility is the postponement of pregnancy: adverse effects of increasing age on reproductive capacity include decreased conception rates and increased pregnancy losses. The increasing incidence of pelvic inflammatory disease is also thought to be a major cause: infertility increases from 11% after one episode of infection to 54% after three episodes. Pelvic infections lead to scar formation around the ovaries and the fallopian tubes, thereby impeding the transport of oocytes for fertilization.

Ovulatory dysfunction accounts for approximately 20% of cases of infertility. Ovulation, the release of the oocyte from the ovary, is dependent on the timely secretion of the two gonadotropic hormones from the pituitary gland, follicle-stimulating hormone (FSH) and luteinizing hormone (LH). This process can be monitored by determining basal body temperature or the serum concentration of the hormone progesterone, which is secreted by the ovary after ovulation. Gonadotropic hormone release can be impeded by certain medications and by many disease states, including stress, anorexia nervosa, weight loss, and thyroid disease. Anovulation can be treated by replacing the deficient gonadotropins with human menopausal gonadotropin, a mixture of follicle-stimulating and luteinizing hormones. In addition, the drug clomiphene citrate increases the endogenous secretion of those hormones from the pituitary gland. *See* ENDOCRINE MECHANISMS; PITUITARY GLAND.

Fallopian tube and uterine abnormalities account for 25% of the cases of infertile couples. These abnormalities include defective development as well as scar formation after surgery or infection. The initial evaluation of this portion of the reproductive tract includes a hysterosalpingogram, that is, an x-ray of the uterus and fallopian tubes. After injection of a radiopaque dye through the cervix, the intrauterine contour and patency of the fallopian tubes are revealed. Not uncommonly the region of the fallopian tube closest to the uterus may be obstructed by scar tissue, but the obstruction can be microscopically removed or the fallopian tube can be surgically cut and reconnected. Occasionally, this area is closed by mucous plugs or tubal spasm. This condition can be successfully treated by directing a fine catheter into the fallopian tube to relieve the spasm or dislodge the plugs. This approach eliminates the need for microscopic surgery and yields comparable pregnancy rates. Uterine defects can be resected, but sometimes several attempts are necessary.

In endometriosis, tissue that normally lines the uterus is found on the pelvic lining. Severe forms of endometriosis involve the ovaries and fallopian tubes, which doubles the likelihood of infertility in those women. Identification of endometriosis is made at surgery, and the condition is treated either by surgical excision of the aberrant tissue or by subsequent medical treatment.

The cervix is more than a conduit for the passage of sperm to the uterus. Sperm may reside in cervical crypts for as long as 48 h, possibly with some enhancement of function. Abnormalities in cervical mucus production may result from trauma, surgery, and diethylstilbestrol (DES) exposure before birth. The cervical environment may be evaluated by examining the mucus 4–6 h after intercourse for the presence of viable sperm. The treatment of cervical abnormalities remains controversial, but may include estrogen; or another approach is to bypass this obstruction by using intrauterine transfer of semen (artificial insemination).

The male. Male infertility resulting from abnormal semen may be due to developmental defects, genitourinary infections, or varicocele. An evaluation of sperm count, motility, and morphology is helpful, but in most cases the cause of the abnormality remains undetermined. No treatment other than donor insemination is available for these couples.

Unexplained infertility. In approximately 10% of infertile couples, a thorough evaluation reveals no cause, and unexplained infertility is diagnosed. In most of these cases, diagnostic laparoscopy is recommended, involving direct visualization of the woman's pelvis through a laparoscope inserted through the abdominal wall. In nearly 50% of women with unexplained infertility, pelvic abnormalities are found and are most often caused by endometriosis and scar formation.

Assisted reproduction. In vitro fertilization takes place through the controlled hyperstimulation of ovulation followed by surgical extraction of mature oocyte from the ovaries. The oocytes are allowed to incubate in a petri dish with the semen specimen. After the oocytes are fertilized, the embryos are allowed to mature and are then transferred into the uterus. In vitro fertilization may be appropriate when unexplained infertility, nonfunctioning fallopian tubes, endometriosis, or decreased sperm counts prevent conception; several thousand children have been conceived by using such a procedure.

Another method of assisted reproduction is gamete intrafallopian tube transfer (GIFT). Ovulation is stimulated, as with in vitro fertilization, but the extracted oocytes are placed directly into the fallopian tube with the semen specimen. Incubation occurs within the fallopian tube rather than outside the body. The procedure is undertaken for the same reasons as those for in vitro fertilization; however, in this case the fallopian tubes must be normal. In micromanipulation, microscopic probes hold oocytes and embryos, so that procedures to improve fertilization or genetic diagnostic tests on embryos can be performed. With severely low sperm numbers, even in vitro fertilization is generally not a viable option. Intracytoplasmic sperm injection can result in a viable pregnancy when a single sperm is injected into the oocyte by micromanipulation. This approach offers an option to childless couples, in addition to donor insemination. Also, micromanipulation allows sophisticated chromosomal analysis of embryos. *See* FERTILIZATION (ANIMAL); PREGNANCY; REPRODUCTIVE SYSTEM; REPRODUCTIVE SYSTEM DISORDERS.

Richard E. Leach; Roger D. Kempers

Bibliography. S. J. Behrman et al. (eds.), *Progress in Infertility*, 4th ed., 1993; J. Cohen et al., *Micromanipulation of Human Gametes and Embryos*, 1992; D. Mishell, *Infertility, Contraception and Reproductive Endocrinology*, 3d ed., 1991; G. Sher, V. A. Marriage, and J. Stoess, *In Vitro Fertilization*, 1989.

Infinity

The terms infinity and infinite have a variety of related meanings in mathematics. The adjective finite means "having an end," so infinity may be used to refer to something having no end. In order to give a precise definition, the mathematical domain of discourse must be specified.

Infinity in sets. A simple and basic example of an infinite collection is the class of natural numbers, or positive integers. A fundamental property of positive integers is that after each integer there follows a next one, so that there is no last integer. Now it is necessary in mathematics to treat the collection of all positive integers as an entity, and this entity is the simplest infinity, or infinite collection.

Suppose that two collections A and B of objects can be set into one-to-one correspondence; that is, each object in A is paired with one and only one object in B, and each object in B belongs to just one of these pairs. Then A and B are said to contain "the same number" of objects, or to have "the same cardinal number." This is equivalent to the ordinary meaning of "the same number," obtained by counting, when the collections A and B are finite, and is taken as the definition of this phrase in the general case.

Now let A consist of all the positive integers, and let B consist of the even integers. Then A and B have the same number of elements, since each element of A can be paired with its double in B; thus elements can be removed from an infinite collection without reducing the number of its elements. This property distinguishes infinite collections from finite ones, and indicates that the statement "the number of elements of a collection A is greater than the number of elements of a collection B" should be taken to mean "B can be put into one-to-one correspondence with part of A, but not with the whole of A."

Consider next the number of subsets of a given collection A. For example, let $A = \{a, b, c\}$; that is, A consists of the three letters a, b, and c. Then the sets $\{a\}$, $\{b\}$, $\{c\}$, $\{a, b\}$, $\{a, c\}$, $\{b, c\}$, are subsets of A. To these, for convenience, can be adjoined the null set, containing no elements, and the set A itself. This makes a total of $8 = 2^3$ subsets. It is readily proved that a set A having a finite number n of elements has 2^n subsets. Thus the number of subsets of a finite set is always greater than the number of elements. By a different method, the same property may be proved to hold also for infinite sets. Hence for every infinite number there is a greater number. In other words, there exist infinitely many distinct infinities. Sets having the same number of elements as the collection of positive integers are called countably infinite, and this is the smallest infinity.

Infinity in limits of functions. The term infinity appears in mathematics in a different sense in connection with limits of functions. For example, consider the function defined by $y = 1/x$. When x tends to 0, y approaches infinity, and the expression may be written as

$$\lim_{x\to 0} y = \infty$$

Precisely, this means that for an arbitrary number $a > 0$, there exists a number $b > 0$ such that when $0 < x < b$, then $y > a$, and when $-b < x < 0$, then $y < -a$. This example indicates that it is sometimes useful to distinguish $+\infty$ and $-\infty$. The points $+\infty$ and $-\infty$ are pictured at the two ends of the y axis, a line which has no ends in the proper sense of euclidean geometry.

Infinity in geometry. In geometry of two or more dimensions, it is sometimes said that two parallel lines meet at infinity. This leads to the conception of just one point at infinity on each set of parallel lines and of a line at infinity on each set of parallel planes. With such agreements, parts of euclidean geometry can be discussed in the terms of projective geometry. For example, one may speak of the asymptotes of a hyperbola as being tangent to the hyperbola at infinity. Note that the points at infinity which are adjoined to a euclidean line or plane are chosen in a manner dictated by convenience for the theory being discussed. Thus only one point at infinity is adjoined to the complex plane used for geometric representation in connection with the theory of functions of a complex variable.

Other concepts. Other types of infinities may be distinguished when properties other than the mere cardinal number of a set are being considered. For example, a line and a plane contain the same number of points, but when continuity considerations are important the line is said to be made up of a single infinity of points, whereas the plane has a double infinity, or ∞^2, points. This is because there exists a one-to-one continuous correspondence between the line and a subset of the plane, but there does not exist such a correspondence between the whole plane and the line or any part of it. As another example, an infinite set may be ordered in different ways so as to have different ordinal numbers. *See* CALCULUS.

Lawrence M. Graves

Bibliography. R. G. Bartle and D. R. Sherbert, *Introduction to Real Analysis*, 3d ed., 1999; W. Rudin, *Real and Complex Analysis*, 3d ed., 1991.

Inflammation

The local response to injury, involving small blood vessels, the cells circulating within these vessels, and nearby connective tissue.

The early phases of the inflammatory response are stereotyped: A similar sequence of events occurs in a variety of tissue sites in response to a diversity of injuries. The response characteristically begins with hyperemia, edema, and adherence of the circulating white blood cells to endothelial cells. The white cells then migrate between the endothelial cells of the blood vessel into the tissue. The subsequent development of the inflammatory process is determined by factors such as type and location of injury, immune state of the host, and the use of therapeutic agents.

Cardinal signs. The discomfort of inflammation may be attendant on a sunburn, a mosquito bite, a cut, an abscess, or a vaccination, and so it is not surprising that historical records of inflammation date to the earliest known medical writings, in the Egyptian papyri. It remained, however, for Celsus (about 30 B.C.–A.D. 38), a Roman, to enumerate the four cardinal signs of inflammation: redness, swelling, heat, and pain. Redness is caused by hyperemia, that is, an increased number of red blood cells in the local capillary bed. Heat results from increased flow of blood in the small blood vessels, and the edema, or swelling, represents the accumulation of extracellular fluid in the connective tissue. Stimulation of nerve endings by agents released during the inflammatory process causes pain. *See* EDEMA.

While it is possible to observe the events of inflammation in human skin, a detailed study of the dynamic cellular events requires a more convenient site in an experimental animal. Tissues such as rat mesentery, frog tongue, hamster cheek pouch, or bat wing have the virtues of accessibility, vascularity, and transparency necessary for microscopic study of the inflammatory process.

Vascular changes. Following a mild injury, there is a fleeting constriction of the smallest arterial branches (arterioles) in the viable tissue close to the injury, lasting from 5 s to about 5 min, followed by dilation of the same arterioles. That leads to engorgement and dilation of the capillaries, a change that extends over an area larger than the constriction site. Constriction of the smallest veins (venules) may also contribute to capillary engorgement.

Along with those blood vessel changes, the steady-state flow of water back and forth across the vessel walls is disrupted; when the outflow of water from the blood vessel exceeds the return, extravascular extracellular fluid collects. An excess of fluid in the extracellular compartment of the connective tissue is termed edema. The vessel walls also become more permeable to the large protein molecules that circulate in the plasma; those molecules leak into the tissue and induce the outward movement of water. Blood flow is slowed in the immediate vicinity of the injury by constriction of the venules and concentration of the blood cells. The latter phenomenon is a result of loss of water from the circulation. Cessation of blood flow (stasis) may occur and be associated with formation of a thrombus, a blood clot in the small vessels, or may be only transitory with prompt restoration of flow. *See* CIRCULATION; THROMBOSIS.

The leukocytes, at first predominantly neutrophilic granulocytes, adhere to endothelial cells that line the capillaries and venules, and then migrate through the vessel wall into the edematous connective tissue, drawn there by chemotactic agents that are released at the site of the injury.

If the tissue injury breaks small blood vessels, another mechanism comes into play. Exposure of circulating blood to collagen leads to clumping of platelets, and the resulting aggregate forms a temporary hemostatic plug that prevents free bleeding from the broken vessel. The clotting mechanism is set in action somewhat more slowly, and the fibrin formed thereby bolsters the plug. *See* COLLAGEN.

Cellular changes. In contrast to the vascular response, the cellular response in inflammation is varied and serves to characterize the different types of inflammation. Participating cells come from two sources, the circulating blood and the local connective tissue. The circulating leukocytes are divided into six distinguishable cell types: granulocytes (eosinophil, neutrophil, and basophil), small and large lymphocytes, and monocytes. Fibroblasts and mast cells are solely connective-tissue cells. Macrophages and giant cells arise in the inflammatory locus from monocytes. *See* BLOOD.

Neutrophils are normally the most abundant leukocytes in the circulation and are the first cells to accumulate in an inflammatory site. They have the capacity to ingest and kill bacteria; the cytoplasmic granules, which identify the neutrophil, contain the enzymes responsible for killing and digesting the bacteria. When the number of circulating neutrophils is greatly decreased, as in patients after treatment with certain drugs or exposure to nuclear radiation, frequent and severe infections may occur. In spite of antibiotic therapy, such patients often succumb to infection. *See* PHAGOCYTOSIS.

Eosinophils are much rarer than neutrophils but are also phagocytic and may be considerably increased in patients with allergies. The basophil, least abundant of the granulocytes, is similar to the tissue mast cell; both store and release histamine and are thus responsible for hives and the symptoms of hay fever. There is evidence that the beneficial effects of both eosinophils and basophils are affected in dealing with multicellular parasites rather than bacteria or viruses. Lymphocytes are second to neutrophils in abundance and are very important in immune responses, including the rejection of grafts of foreign tissues, and the attack of donor cells. Monocytes, like lymphocytes, lack specific granules. This cell type when stimulated has the potential to change into a macrophage, which, like a neutrophil, can ingest and kill bacteria. Fibroblasts are responsible for synthesis of collagen and other components of the extracellular connective tissue during the healing process. *See* HISTAMINE.

Cause-effect relationship. An inflammatory response may be induced in a variety of ways. Causes of inflammation include trauma, heat, ultraviolet light, x-rays, bacteria, viruses, many chemicals such as turpentine and croton oil, and certain inherently innocuous but foreign substances (antigens) which evoke immune responses. Although many of the components of inflammation are common to a variety of inflammatory responses, particularly in the early stages, there is considerable diversity in fully developed responses; the response to an abscess is very different from that of a burn, for instance. The character of the injury, its severity, and the site of injury modify the progress of the inflammatory response, as does therapeutic intervention. *See* ANTIGEN; IMMUNITY.

A local inflammatory response is usually accompanied by systemic changes: fever, malaise, an increase in circulating leukocytes (leukocytosis), and increases in specific circulating proteins called acute-phase reactants. Such signals and symptoms are often helpful to the physician, first as clues to the presence of inflammation and later as an indication of its course.

Mediators of inflammation. The process of inflammation, both vascular and cellular, is orchestrated by an array of molecules produced locally. They variously induce the adhesion of circulating leukocytes to the endothelium, increase small blood vessel permeability to proteins, increase blood flow, direct the migration of inflammatory cells, enhance phagocytic activity, incite proliferation and differentiation of inflammatory cell types locally, and elicit the systemic correlates of inflammation. These mediators include histamine, leukotrienes, prostaglandins, complement components, kinins, antibodies, and interleukins. Many anti-inflammatory drugs function by preventing the formation of those mediators or by blocking their actions on the target cells whose behavior is modified by the mediators.

Types. Inflammation is frequently described in terms of its time course. Acute inflammation develops rapidly, in a matter of hours to days, and is of relatively short duration. In acute inflammation the neutrophil is the predominant cell type, and hyperemia and edema are prominent. Chronic inflammation develops over a period of weeks to months and is characterized by an infiltrate of lymphocytes, monocytes, and plasma cells—the chief antibody-producing cells. Local differentiation and proliferation of macrophages is characteristic of chronic inflammation.

Granulomatous inflammation is a specific type of chronic inflammation in which a discrete nodular lesion (granuloma) is formed of macrophages, lymphocytes, plasma cells, and giant cells arranged around a central mass of noncellular material. Granulomas, typical of tuberculosis and fungus infection, also occur in rheumatoid arthritis. *See* GRANULOMA INGUINALE; TUBERCULOSIS.

Abscesses and cellulitis are specific forms of acute inflammation; the former term denotes a localized collection of pus composed of necrotic debris derived from dead tissue cells and neutrophils. Cellulitis is characterized by diffuse hyperemia and edema with an extensive neutrophil infiltrate, and often little tissue destruction. When inflammation causes erosion of an epithelial surface, the lesion is termed an ulcer. *See* ULCER.

Function. Inflammation is basically a protective mechanism. The leakage of water and protein into the injured area brings humoral factors, including antibodies, into the locale and may serve to dilute soluble toxic substances and wash them away. The adherence and migration of leukocytes brings them to the local site to deal with infectious agents. There are also instances in which no causative toxic substance or infectious agent can be found to account for the inflammation. This is the case in rheumatoid arthritis and rheumatic fever. Such diseases may be examples in which an uncontrolled or misdirected inflammatory response with an autoimmune component is turned against the host. *See* ARTHRITIS; AUTOIMMUNITY; INFECTION; RHEUMATIC FEVER. David Lagunoff

Bibliography. I. L. Bonta and M. A. Bray (eds.), *The Pharmacology of Inflammation*, 1985; J. I. Gallin et al. (eds.), *Inflammation: Basic Principles and Clinical Correlates*, 3d ed., 1999; G. P. Lewis, *Mediator in Inflammation*, 1986; J. G. Lombardino (ed.), *Nonsteroidal Antiinflammatory Drugs*, 1985; P. Venge and A. Lindbom, *Inflammation: Basic Mechanisms, Tissue Injuring Principles, and Clinical Models*, 1985.

Inflammatory bowel disease

Inflammatory bowel disease is a general term for two closely related conditions, ulcerative colitis and regional enteritis or Crohn's disease. The diseases can affect the colon, distal, small intestine, and sometimes other portions of the gastrointestinal tract as well as several sites outside the gastrointestinal tract. In 15–25% of cases limited to the colon, ulcerative colitis and Crohn's disease cannot be distinguished by clinical manifestations, x-ray examination, or even pathology. For this reason the broad term inflammatory bowel diseases is useful. The cause of these diseases is unknown.

Ulcerative colitis. Ulcerative colitis, an inflammatory condition limited to the colon, primarily affects the mucosa or lining of the colon. Marked inflammation gives rise to small ulcerations and microscopic abscesses that produce bleeding. The condition tends to be chronic, alternating between periods of complete remission and episodes of active and even life-threatening disease. Colitis may involve a varying extent of the colon; when limited to the rectum, it is known as ulcerative proctitis; it may be confined to the descending colon; or it may affect the entire colon, being known in this case as universal colitis.

The most common symptoms include rectal bleeding and diarrhea; blood is almost always present during a flare-up. Occasionally, there may be signs of toxicity, with other constitutional symptoms. Manifestations of ulcerative colitis, as well as Crohn's disease, outside the gastrointestinal tract include eye and joint inflammation and skin disorders. Extraintestinal disease occurs in 15–20% of patients with ulcerative colitis.

The greatest concern in long-term management is the risk of colon cancer; ulcerative colitis must therefore be monitored as a precancerous condition. When limited to the rectum and sigmoid, the disease does not seem to increase the risk of colon cancer, but when it is present beyond those most distal segments, the risk begins after about 10 years of disease and progresses steadily. The risk of colon cancer can be evaluated by microscopic examination of the epithelial cells of the colon lining. Biopsies can be examined for cell changes known as dysplasia or precancer. If there is no dysplasia on multiple samples, the risk of cancer remains low; if dysplasia is found, the risk is high.

Ulcerative colitis is treated with corticosteroids and a salicylate–sulfa drug combination known as sulfasalazine. If medication is not effective, surgical removal of the colon eliminates the disease but necessitates the creation of an opening in the abdominal wall from which the contents of the intestine can pass. The procedure, known as an ileostomy, redirects the lower end of the small intestine through the surface of the body where a collection appliance is attached to the skin. Alternatives to this method include an internal pouch or a valve mechanism. Other surgical techniques modify the distal rectum or anus so that evacuation can take place through the normal anatomic site.

Crohn's disease. Crohn's disease, also known as regional enteritis, granulomatous colitis, and terminal ileitis, affects the colon and small intestine, and rarely the stomach or esophagus. Like ulcerative colitis, it is chronic and of unknown etiology. The two diseases have many similarities, including their treatment. Pathologically, however, the findings are usually distinct. In Crohn's disease, chronic inflammation is present and is usually accompanied by granulomas. The inflammation involves the full thickness of the intestinal wall, often with bowel narrowing and obstruction of the lumen. Abdominal cramps, alteration of bowel function, and diminished food intake are common.

In the young, retardation of growth and of sexual maturation are frequently observed. Crohn's disease involving different parts of the gastrointestinal tract produces symptoms related to those regions. For example, extensive disease of the distal small intestine can interfere with absorption of nutrients, particularly vitamin B_{12} and bile salts, which are absorbed somewhat specifically in that region. Deficiency of bile salts can lead to malabsorption of fats.

Corticosteroid medications are utilized to treat the condition. Sulfapyridine used in ulcerative colitis can be effective, particularly when Crohn's disease affects the colon. Metronidazole is effective for treating inflammation in and around the anal region, which is more common in Crohn's disease than in ulcerative colitis. Surgery may be required in cases of bowel obstruction, hemorrhage, or severe debility. Although Crohn's disease carries a slight increase in the risk of cancer, the risk is modest when compared to ulcerative colitis. *See* DIGESTIVE SYSTEM; GASTROINTESTINAL TRACT DISORDERS.

Leonard A. Katz

Bibliography. A. Anagostides, H. J. Hodgson, and J. B. Kirsner (eds.), *Inflammatory Bowel Disease*, 1991; S. B. Hanauer and J. B. Kirsner (eds.), *Inflammatory Bowel Disease*: *A Guide for Patients and Their Families*, 1985; D. Rachmilewitz (ed.), *Inflammatory Bowel Diseases*, 1986.

Inflationary universe cosmology

A theory of the evolution of the early universe, motivated by considerations from elementary particle physics as well as certain paradoxes of standard big bang cosmology, which asserts that at some early time the observable universe underwent a period of exponential, or otherwise superluminal, expansion.

In order to resolve the various paradoxes of the big bang theory, inflationary theories predictthat during this inflationary epoch the scale of the universe increased by at least 28 orders of magnitude. The period of superluminal expansion (inflation) in most cases is caused by the appearance of a nonzero constant energy density in the universe associated with the postulated existence of a phase transition between different ground-state configurations of matter that occurs as the universe expands and cools. After the transition is completed, the constant energy density is converted into the energy density of a gas of relativistic particles. At this point, inflationary scenarios match the standard big bang cosmological model.

Origin of inflationary models. The suggestion of an inflationary period during the early universe, in connection with a specific model, was first made in 1980 by A. Guth. (Somewhat earlier a less concrete but not entirely unrelated possibility was discussed by A. A. Starobinskii.) Guth proposed, based on a consideration of recently proposed grand unification theories in particle physics, that phase transitions—associated with the breaking of certain symmetries of the dynamics governing the interactions of matter—could occur at sufficiently high temperatures and have important consequences in the early universe.

Symmetry breaking and phase transitions. The concept of symmetry is of fundamental importance in physics. In particular, a symmetry can be associated with each of the known forces in nature, and each symmetry is, in turn, associated with the existence of certain conserved quantities, such as electric charge in the case of electromagnetism. Surprisingly, a physical system in its ground state may not possess a symmetry that could be fundamental to the basic physics governing its dynamics. The most familiar example is a ferromagnet made up of many individual elementary magnetic dipoles. The equations of electromagnetism are manifestly rotationally invariant: there is no intrinsic north or south. However, the ground state of such a spin system will involve all the spins aligned in one particular direction, yielding the familiar case of a permanent magnet. In this case, it is said that the rotational symmetry of the equations of electromagnetism has been spontaneously broken. The signal for this is that the average value of the total spin of the system points in a certain direction. If the system is heated, so that all the spins become excited and randomly oriented, the net spin of the system approaches zero, its magnetic field vanishes, and rotational symmetry is restored. The total spin of the system is referred to as an order parameter, because its value tells which state the system is in, and a change in the order parameter is an indication of the presence of a phase transition between different ground states of the system. *See* PHASE TRANSITIONS; SYMMETRY BREAKING; SYMMETRY LAWS (PHYSICS).

Phase transitions in particle physics. A situation similar to the case described above occurs in particle physics. At the temperatures and energies that occur in the universe now, there isavast difference in the nature of the known fundamental forces. For two of these forces—the electromagnetic force, and the so-called weak force, which governs the beta decay processes important to energy production in the Sun—it is now known that the perceived differences are the result of spontaneous symmetry breaking. Above a certain energy or temperature, the weak and electromagnetic interactions appear exactly the same. However, at a critical temperature, the ground state of matter, which in particle physics is called the vacuum state, breaks the symmetry relating the two interactions, and below this temperature they appear quite different. The signal for this symmetry breaking is again the appearance of a nonzero value in a certain order parameter of the system; in this case it is referred to as a vacuum expectation value. In the case of the ferromagnet, the order parameter was a spin, and the ground state after symmetry breaking had a net magnetization. In the case of the particle physics system, the order parameter describes the ground state expectation value of a certain elementary particle field, and the vacuum state after symmetry breaking has a net charge that would have been zero had the symmetry not been broken.

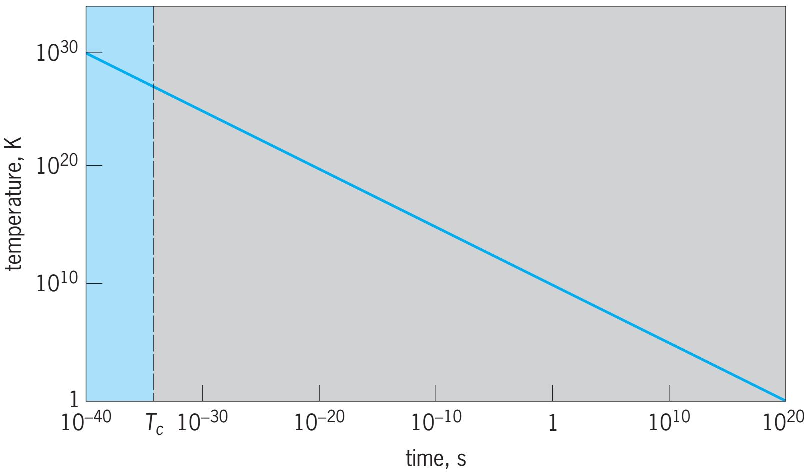


Fig. 1. **Temperature versus time after beginning of expansion for the standard big bang model. Before T_c, inflationary scenarios differ from the standard model.**

In particle physics, if a symmetry is broken at a certain scale of energy, particles that transmit the force of nature associated with that symmetry can have masses characteristic of this energy scale. This is indeed what happens in the case of the weak interactions, which are weak precisely because the particles that transmit the weak force are so massive. However, once the temperature of a system is so great that the energy it takes to produce such a massive particle is readily available in thermal energy, the distinction between this force and electromagnetism, which is transmitted by massless particles (photons), disappears and the two forces become unified into one.

Shortly after it was recognized that these two forces could be unified into one, it was proposed that all three observed forces in nature outside of gravity might be combined into one grand unified theory. There are reasons to believe that the scale of energy at which all the forces would begin to appear the same is truly astronomical—some 16 orders of magnitude greater than the mass of the proton. Such energies are not achieved in any terrestrial environment, even in the laboratory. The only time when such a scale of energy was common was in the earliest moments of the big bang fireball explosion. *See* ELEMENTARY PARTICLE; FUNDAMENTAL INTERACTIONS; GRAND UNIFICATION THEORIES.

“Old” inflationary model. Based on these facts, Guth reasoned that if grand unification symmetries are indeed broken at some large energy scale, then a phase transition could occur in the early universe as the temperature cooled below the critical temperature where symmetry breaking occurs. According to the standard big bang model of expansion, the time at which this would occur would be about 10^{-35} s after the initial expansion had begun (**Fig. 1**). This time is far earlier than the times previously considered in cosmological analyses using the big bang model. For instance, the time at which it is predicted that many of the light elements were first formed in the process of primordial nucleosynthesis is of the order of 1 s, some 35 orders of magnitude later in time.

First-order phase transitions. As Guth demonstrated, the effects of such a phase transition in the early universe could be profound. In order to calculate the dynamics of a phase transition, it is necessary to follow the behavior of the relevant order parameter for the transition. This is done by determining the energy (actually the thermodynamic free energy in the case of systems at finite temperatures) of a system as a function of the order parameter, and following the changes in this energy as the temperature changes. Typically (**Fig. 2**), at some high temperature T, the minimum of the relevant energy function is at zero value of the order parameter, the vacuum expectation value of a certain field. Thus the ground state of the system, which occurs when this energy is a minimum, will be the symmetric ground state. This case is analogous to the spin system, which at high temperatures is disordered, so that no preferred direction is picked out. As the temperature is decreased, however, at a certain critical temperature T_c a new minimum of the energy appears at a nonzero value of the order parameter. This is the symmetry-breaking ground state of the system. *See* FREE ENERGY.

How the system makes a transition between the original, symmetric ground state and the new ground state depends on the shape of the energy curve as a function of the order parameter. If there is a barrier between the two minima (Fig. 2), then classically the system cannot make a transition between the two states. However, it is a well-known property of quantum mechanics that the system can, with a certain very small probability, tunnel through the barrier and arrive in the new phase. Such a transition is called a first-order phase transition. Because the probability of such a tunneling process is small, the system can remain for a long time in the symmetric phase before the transition occurs. This phenomenon is called supercooling. When the transition finally begins, “bubbles” of new phase locally appear throughout the original phase. As more and more bubbles form, and the original bubbles grow (in the case of such a phase transition in particle physics the bubbles grow at the speed of light), eventually they combine and coalesce until all of the system is in the new phase. Another example of this type of phenomenon is the case of a supercooled liquid such as water that has been continually stirred as it has been cooled. At a certain point, ice crystals will spontaneously form throughout the volume of water, and grow until all of the water has turned to ice. *See* NONRELATIVISTIC QUANTUM THEORY.

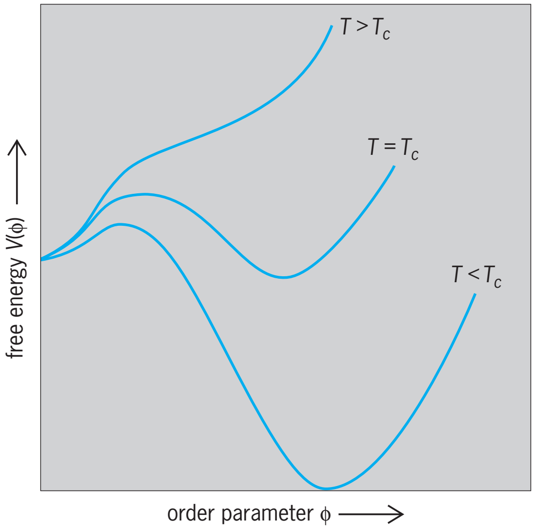


Fig. 2. Free energy as a function of the order parameter for the case of a first-order transition that is typical of symmetry breaking in grand unification theories.

Inflation. Such a phase transition, when treated in the context of an expanding universe, can result in remarkable behavior. The dynamics of an expanding universe are governed by Einstein’s equations from the theory of general relativity. When these equations are solved by using the properties of ordinary matter, it is seen that the rate of expansion of the universe, characterized by the Hubble parameter H—which gives the velocity of separation of objects—slows down over time in a way that is dependent on the temperature of the universe, T. (Specifically, in a radiation-dominated universe, H is dependent on T^2.) This slowdown occurs because the energy density of matter, which is driving the expansion, is decreased as the matter becomes more dilute because of the effects of expansion.

However, once the temperature is below the critical temperature for the transition (Fig. 2), the metastable symmetric phase has a higher energy than the new lower-energy symmetry-breaking phase. Until the transition occurs, this means that the symmetric phase has associated with it a large constant energy density, independent of temperature. When this constant energy density is placed on the right-hand side of Einstein’s equations, where the energy density of matter appears, it is found that the resultant Hubble parameter describing expansion is a constant. Mathematically, this implies that the scale size of the universe increases exponentially during this supercooling phase. This rapid expansion is what is referred to as inflation. Once the phase transition is completed, the constant energy density of the original phase is converted into the energy density of normal matter in the new phase. This energy density decreases with expansion, so that the universe reverts to the slower rate of expansion characteristic of the standard big bang model. *See* HUBBLE CONSTANT; RELATIVITY.

Successes of inflation. Guth pointed out that a period of exponential expansion at some very early time could solve a number of outstanding paradoxes associated with standard big bang cosmology.

Flatness problem. In 1979, it was pointed out that present observations seem to imply that either the present era is a unique time in the big bang expansion or the initial conditions of expansion had to be fine-tuned to an incredible degree. The observation involves the question of whether the universe is now open, closed, or flat. A closed universe is one where there is sufficient mass in the universe to eventually halt and reverse the observed expansion because of net gravitational attraction; an open universe will go on expanding at a finite rate forever; and a flat universe forms the boundary between these two cases, where the expansion rate will continually slow down and approach zero asymptotically. Measurements of the observed expansion rate, combined with measurements of the observed mass density of the universe, yield a value for the density parameter Ω that rapidly approaches 0 for an open universe, infinity for a closed universe, and is exactly equal to 1 for a flat universe. All measurements of Ω yield values between about 0.1 and 2. What is so strange about this measurement is that theory suggests that once the value of Ω deviates even slightly from 1, it very quickly approaches its asymptotic value far away from 1 for open or closed universes. Thus, it is difficult to understand why, after 10^{10} years of expansion, the value of Ω is now so close to 1. At a quantitative level, the problem is even more mysterious. In order to have its present value in the range given above, at the time of nucleosynthesis the value of Ω would have had to have been equal to 1 within 1 part in 10^{15}.

Inflation naturally explains why Ω should exactly equal 1 in the observable universe today. Einstein's equations for an expanding universe can in the cases of interest be written in the form of the equation below, where t is the time after beginning of expansion, R is the cosmic scale factor, related to the size of the universe in closed universe models, ρ is the energy density of matter, and K is a constant that is equal to 0 for flat universe models.

$$1 + \frac{K}{R(t)^2 \rho(t)} = \Omega(t)$$

Normally, for an expanding universe, the energy density of matter decreases with scale at least as fast as $1/R^3$, so that the left-hand side deviates from 1 as time goes on. This is the quantitative origin of the flatness problem. However, in an inflationary phase, $\rho(t)$ remains constant, while R increases exponentially. Thus, during inflation $\Omega(t)$ is driven arbitrarily close to 1 within the inflated region. If there are some 28 orders of magnitude of exponential expansion, which is possible if supercooling lasts for some time, then $\Omega(t)$ need not have been finely tuned to be close to 1, even if inflation occurred as early as the Planck time (10^{-45} s).

Horizon problem. An equally puzzling problem that inflationary cosmology circumvents has to do with the observed large-scale uniformity of the universe. On the largest observable scales, the universe appears to be largely isotropic and homogeneous. In particular, the 3-K microwave radiation background, which in different directions has propagated from distances separated by 10^{10} light-years, is known to be uniform in temperature to about 1 part in 100,000. *See* COSMIC BACKGROUND RADIATION.

This observed uniformity may not seem so puzzling, until an attempt is made to derive it in the context of any set of reasonable initial conditions for the big bang expansion. Physical effects can propagate at best at the speed of light. The distance a light ray could have traveled since the big bang explosion, which is also thus the farthest distance at which one object can affect another, is called the horizon. The horizon size increases linearly with time, since light travels with a constant velocity even in an expanding universe. Thus, the size of the horizon today, which is about 10^{10} light-years, is much larger than the horizon size at the time the radiation in the microwave background was emitted. In particular, in the standard big bang model the sources of the radiation observed coming from opposite directions in the sky were separated by more than 90 times the horizon distance at the time of emission. Since these regions could not possibly have been in physical contact, it is difficult to see why the temperature at the time of emission was so uniform in all directions.

Even if some very isotropic initial conditions are postulated for the big bang expansion to account for this uniformity, a quantitative problem similar to the flatness problem is encountered in attempting to account for the degree of uniformity now observed. Small fluctuations in energy density tend to grow as the universe evolves because of gravitational clumping. Indeed, observed clumping on the scale of galaxies and smaller attests both to this fact and to the existence of some initial irregularities in the distribution of matter. However, since irregularities grow but the universe is relatively smooth on large scales, any initial conditions for the big bang expansion imposed at early times, say 10^{-45} s, would have to involve an absurdly uniform distribution of matter on the largest scales.

Inflation solves the horizon problem very simply. If the observed universe expanded by 28 orders of magnitude in a short period at a very early time, then in inflationary cosmology it originated from a region 10^{28} times smaller than the comparable region in the standard big bang model extrapolated back beyond that time. This makes it quite possible that at early times the entire observed universe was contained in a single horizon volume (**Fig. 3**).

Problems with "old" inflation. The list of cosmological problems resolved by the original inflationary universe model is quite impressive. Moreover, its origin in current elementary-particle-physics ideas made it more than an artificial device designed specifically to avoid cosmological paradoxes. Unfortunately, the original inflationary scenario was fundamentally flawed. The central problem for this scenario was how to complete the phase transition in a uniform way. The phase transition began by the formation of bubbles of one phase nucleating amidst the initial metastable phase of matter. While these bubbles grow at the speed of light once they form, the space between the bubbles is expanding exponentially. Thus, it is extremely difficult for bubbles to

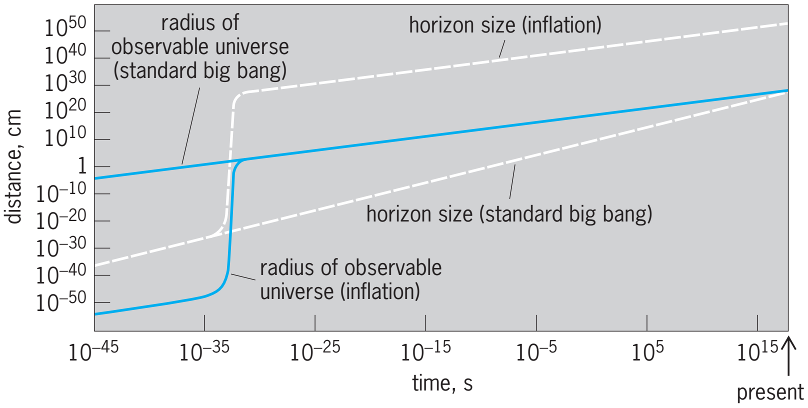


Fig. 3. Horizon size and radius of the observable universe as functions of time after beginning of expansion, in the standard big bang and inflationary models. (*After A. H. Guth and P. J. Steinhardt, The inflationary universe, Sci. Amer., 250*(5)*:116–128, May 1984*)

eventually occupy all of space, as is required for the completion of the transition. Moreover, even if bubbles are forming at a constant rate, each region of space quickly becomes dominated by the largest bubble, which formed earliest. Collisions with much smaller bubbles in this region will not adequately or uniformly dissipate the energy of this largest bubble, so that even if the transition does manage to percolate through space, the final state will be very nonuniform on all scales. Finally, topologically stable objects such as magnetic monopoles and domain walls can form at the intersection of bubbles, and end up with densities after inflation well above those allowed by observation. *See* MAGNETIC MONOPOLES.

While these problems motivated the development of a "new" inflationary scenario, described below, in 1989 it was pointed out that in certain previously examined special models related to so-called Brans-Dicke models of gravitation, the gravitational constant can itself vary during an inflationary phase transition. In this case, the inflationary expansion rate, while very fast, need not be exponentially fast, so that bubbles growing at the speed of light can eventually percolate throughout all of space, thereby completing the phase transition and ending the inflationary expansion. However, many details of these rather special scenarios, which may still leave large remnant inhomogeneities on the scale where bubbles coalesce, remain to be worked out. *See* GRAVITATION.

"New" inflationary cosmology. In 1981 a new approach was developed by A. D. Linde, and independently by A. Albrecht and Steinhardt, which has since become known as new inflation. They suggested that if the energy function (potential; Fig. 2) were slightly changed, then it might be possible to maintain the successful phenomenology of the old inflationary model while avoiding its problems. In particular, they considered a special form of the potential that is extremely flat at the origin (**Fig. 4**). Most important, such functions have essentially no barrier separating the metastable from the stable phase at low temperature. Now, when the universe cooled down below the critical temperature, the order parameter could continuously increase from zero instead of tunneling discretely to a large nonzero value. As long as the potential is sufficiently flat near the origin, however, it can take a long time before the order parameter approaches its value at the true minimum of the potential (a so-called slow-rollover transition), at which point the region will be in the new phase. During this time the region of interest can again be expanding exponentially because of the large constant energy density that is maintained while the order parameter remains near the origin. Thus, in some sense a single bubble can undergo inflation in this scenario. If the amount of inflation in this region is sufficient, the whole observable universe could have originated inside a single inflating bubble.

This key difference between old and new inflationary cosmology accounts for the ability of the latter to bypass the problems for the former. Because in new inflation the observable universe grew from a single inflating bubble, problems of inhomogeneity and topological defects resulting from bubble percolation and collisions are avoided, at least as far as observational cosmology is concerned. Moreover, the problem of completing the phase transition is naturally avoided. Once the order parameter in the inflating region approaches the steep part of the potential, the large constant energy density of the initial phase is reduced, and inflation slows. This energy density is converted into a form of kinetic energy of the changing order parameter. This motion of the order parameter is reflected in a changing charge of the new ground state of matter. Just as a time-varying electric charge produces electromagnetic radiation, so the variation of the order parameter produces a thermal background of real particles. By the time the order parameter has settled at the new minimum of the potential, all of the original constant energy density of the symmetric phase of matter has been converted into energy of real particles at a finite temperature. (Any particle density that existed before inflation has, of course, been diluted away by the vast expansion of the region during inflation.) The evolution from this point on is that of the standard

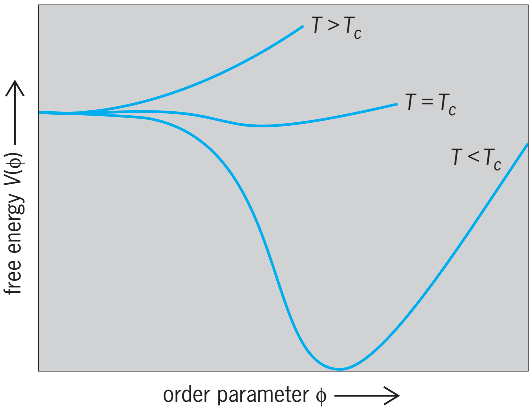


Fig. 4. Free energy as a function of the order parameter for the case of a so-called slow-rollover transition that appears in the new inflationary cosmology.

big bang model, with the initial conditions being a uniform density of matter and radiation at some finite temperature.

Problems with new inflation. New inflation, too, is not without its problems. The type of potential (Fig. 4) needed for a new inflationary scenario is not at all as generic as that postulated in the old inflationary model (Fig. 2). To have such a slow-rollover transition, the parameters of particle physics models must be finely tuned to some degree. Moreover, the temperature to which the universe is reheated after inflation must be large enough that, at the very least, big bang nucleosynthesis can proceed afterward. This again rules out some possible models. No clear candidate model for new inflation has emerged from particle physics.

Another potential problem for any inflationary scenario concerns initial conditions. As discussed above, if an inflationary phase precedes the standard big bang expansion, then it is possible to resolve problems of the standard big bang model related to the unphysical fine tunings that seem necessary at time zero in order for the big bang to evolve into its presently observed form. However, there is also the question of how generic such an inflationary phase is—namely, how special the initial preinflationary conditions must be so that space-time will undergo an inflationary transition in the first place. If these too are unphysical, then inflation may not have really solved any fundamental problems. While Guth suggested in his initial work that the preconditions for inflation were not severe, this view has been questioned by some. Attempts have been made to develop an existence proof that not all initial configurations of the universe can be made isotropic by inflation. This, of course, does not address the fundamental issue of whether such configurations are at all generic.

In 1983, Linde proposed a version of inflation, which he called chaotic inflation, that may in principle address this issue. In particular, he argued that inflation may have occurred very generally in the early universe, even if there was no related phase transition. As long as a scalar field existed with a potential whose scale of energy was sufficiently close to the Planck scale, then, if quantum fluctuations carried the field expectation value far enough away from the origin, the field would relax slowly enough that its energy density would remain nonzero and roughly constant for some time, causing an inflationary phase of expansion. Such a possibility led Linde to suggest that the early universe may have been arbitrarily inhomogeneous. In some regions, inflation may have successfully taken place, even without the spontaneous symmetry breaking associated with a grand unified theory, and in other regions it may not have. He then suggested that the regions in which inflation did take place are in fact most probable. In particular he argued that those values of parameters that allow large amounts of inflation to take place are most probable, if an initial random distribution of regions in the preinflation universe governed by different metric parameters is allowed. In addition, he pointed out that life would form only in those regions that became sufficiently isotropic so that, if many different regions of the universe now exist, it is not surprising that humans live in one that has undergone inflation. This argument, a reformulation of the so-called anthropic principle, is, however, difficult to quantify. In any case, the issue of initial conditions for inflation is the subject of much research, but may require an understanding of quantum gravity for its eventual resolution. *See* CHAOS; QUANTUM GRAVITATION.

Predictions. Whether or not there exists a specific model for new inflationary cosmology, and whether or not understanding inflation may require further refinements in the understanding of quantum gravity, the implications of an inflationary phase in the early universe are profound. As discussed above, many of the fundamental paradoxes of standard big bang cosmology can be resolved. Moreover, it has been demonstrated that new inflationary cosmology naturally allows a derivation from first principles of the spectrum of primordial energy density fluctuations responsible for galaxy formation. Before new inflation, there existed no method, even in principle, for deriving this spectrum for the standard big bang model; it always had to be input as an artificial addition. Remarkably, the spectrum that emerges from inflationary models is a so-called scale-invariant spectrum of perturbations. It is exactly this type of spectrum that had been postulated 10 years before the formulation of the new inflationary model to account for the observed properties of galaxies while maintaining agreement with the observed anisotropy of the microwave background.

In a remarkable discovery, the first observation of anisotropies in the microwave background was announced in April 1992, based on the analysis of 2 years of data on the microwave background structure using the differential microwave radiometer experiment aboard the *Cosmic Background Explorer* (*COBE*) satellite launched by the National Aeronautics and Space Administration (NASA) in 1989. A quadrupole anisotropy in the background was observed at a level of about 5×10^{-6}, just in the range that might be expected for primordial fluctuations from inflation that might also result in the observed distribution of galaxies. Moreover, the correlation of temperature deviations observed across the sky on scales greater than about 10° is remarkably consistent with the scale-invariant spectrum predicted by inflationary models. While neither of these observations conclusively proves the existence of an inflationary phase in the early universe, the fact that they are clearly consistent with such a possibility, and at present with no other scenario, gives great confidence that inflationary models may provide the correct description of the universe in the era preceding the present observed isotropic expansion.

Theoretical and observational factors. Since the early 1990s several factors, both theoretical and observational, have dominated thinking about possible inflationary models.

Hybrid inflation. As the theoretical notion of supersymmetry—a symmetry relating the two different kinds of elementary particles in nature, bosons (integer-spin particles) and fermions (half-integer-spin particles)—has begun to dominate the model building of elementary particle physics, the kind of flat potentials which allow long slow rollover periods of inflation have become more widely used. The problem in such models is not how to have sufficient inflation but rather how to end inflation. A new mechanism, which can be relatively natural in some supersymmetric models, has been proposed in this regard. Called hybrid inflation, this mechanism relies not on the evolution of a single-field order parameter but on several interacting fields. In this case, the evolution of these fields as the universe evolves can be more complex. In certain cases, the preferred, lowest free-energy state of matter can evolve along the flat direction, but as the value of the corresponding field evolves, its couplings to other fields can, at a certain critical value, cause the lowest free-energy state to evolve along a nonflat direction, abruptly ending inflation. The advantages of these models are that inflation can end without fine tuning, primordial density fluctuations that are of an acceptable magnitude can be generated without a great deal of fine tuning, and the models are reasonably well motivated on particle physics grounds. Unfortunately, model building in this area is still sufficiently murky that no obvious set of phenomenologically acceptable models yet exists. *See* SUPERSYMMETRY.

It has also been proposed that the generation of the observed matter-antimatter asymmetry in nature might not occur at temperatures corresponding to the grand unified scale, where the symmetry breaking which motivated Guth's original model is thought to occur, but at much lower energies, corresponding to the scale where the symmetry between the weak and electromagnetic interactions breaks. In this case, it may be possible that following inflation the universe thermalizes at a temperature far below that associated with the original inflationary models. This relaxes some of the constraints on inflationary models, and again favors the possible existence of long flat directions in the potential that might be associated with supersymmetric models. *See* ANTIMATTER; ELECTROWEAK INTERACTION; STANDARD MODEL; WEAK NUCLEAR INTERACTIONS.

Preheating. Original calculations suggested that, following inflation, the universe would quickly thermalize to a temperature comparable to the energy density stored in the original symmetric preinflationary phase of matter. Such a scenario was called reheating. Subsequent calculations have shown that a very different behavior can take place. As the order parameter relaxes about the new minimum of the potential following inflation, its vibrations can release energy in certain nonthermal modes. As a result, immediately following inflation a finite fraction of the released energy might exist in such nonthermal modes. Thus, estimates of particle number densities based on the assumption of thermal equilibrium may be dramatically wrong. Among the possible consequences of this situation, called preheating, is the generation of a large density of topological defects such as cosmic strings, and possibly magnetic monopoles in the postinflationary phase. Since one of the original purposes of inflation was ridding the universe of such defects, it is not yet clear what the implications of this new possibility are for the health of inflationary models. *See* COSMIC STRING.

Dark energy, branes, and a flat universe. Finally, since 1995 it has been increasingly clear, based on measurements of the total clustered mass in the universe in galaxies and clusters of galaxies, that there is not sufficient matter to result in a flat universe today. However, it is now recognized that matter accounts for only part of the total energy density of the universe, with the remainder being associated with some new form of energy, possibly energy stored in the ground state of empty space, called a cosmological constant. This energy may in fact be identical to the energy that is stored during the inflationary phase itself. If this is true, we live in an inflationary universe today. Measurements of the expansion rate of the universe as a function of cosmic time, by observing the recession velocities of certain types of exploding stars, called type 1a supernovae, in very distant galaxies, argue that this is precisely the case, with 30% of the energy of a flat universe in matter and 70% in dark energy. It will probably be necessary to rethink many of the current microphysical notions about the expanding universe to accommodate this newly discovered dark energy. *See* ACCELERATING UNIVERSE; COSMOLOGICAL CONSTANT; DARK ENERGY; SUPERNOVA.

A great deal of theoretical work has been devoted to exploring the possibility that our four-dimensional universe might be embedded in a universe of higher dimensions. In such models—in which our four-dimensional universe, which forms a structure called a brane, can be impacted upon by higher-dimensional physics—the possibility exists that inflation might be caused by new physics associated with other branes embedded in higher-dimensional spaces, instead of being associated with phase transitions due purely to physics on our brane. However, these ideas remain highly speculative. *See* SUPERSTRING THEORY.

Results and prospects. The next-generation advanced cosmic microwave background satellite, the *Wilkinson Microwave Anisotropy Probe* (*WMAP*), has been able to probe many of the fundamental parameters of cosmology, including the Hubble expansion rate, the overall matter density, the spectrum of primordial density fluctuations, and the possible existence of a cosmological constant, to an accuracy of better than 10%. *WMAP* has confirmed many of the generic predictions of inflation, including a flat universe and a roughly constant spectrum of primordial fluctuations. Even more significant perhaps, it achieved sufficient accuracy to determine that the spectrum is not precisely scale-invariant, but changes with wavelength very slightly in a way that is precisely that expected by many inflationary models. (Inflation predicts only a roughly scale-invariant spectrum when detailed estimates are performed. The fluctuations spectrum actually increases slightly with increasing wavelength.) While this is not proof that inflation actually occurred, the consistency between theoretical expectations and observations provides further support for this idea. Moreover, detailed measurements of the spectrum of primordial fluctuations can actually rule out some inflationary models. *See* WILKINSON MICROWAVE ANISOTROPY PROBE.

While all of the current data supports the idea of inflation, because many different inflationary models exist, the question remains whether there is a prediction that can be probed that is unique to inflation. Interest currently is focused on the possibility of measuring a primordial spectrum of gravitational waves, which is also generated during an inflationary phase of the early universe. The spectrum of these waves would also be close to scale-invariant, and the amplitude of the waves would depend upon the energy scale at which inflation occurred. Such a spectrum might produce a measurable polarization in the cosmic microwave background at a level that might be probed by *Planck*, the next-generation cosmic microwave background satellite, or by future ground-based detectors. A positive signal could provide very strong evidence for inflation, but unfortunately the absence of a signal would not provide definitive evidence against it. *See* GRAVITATIONAL RADIATION.

For the moment, inflation remains a beautiful and well-motivated theoretical possibility, which is now more strongly supported by all existing data, and in fact is the only such idea that is consistent with all existing cosmological observations. Hopefully, theoretical and observational developments will help settle the question of whether or not it actually occurred. *See* BIG BANG THEORY; COSMOLOGY; UNIVERSE.

Lawrence M. Krauss

Bibliography. C. L. Bennett, M. S. Turner, and M. White, The cosmic Rosetta stone, *Phys. Today*, 50(11):32–38, November 1997; A. H. Guth, Inflationary universe: A possible solution to the horizon and flatness problems, *Phys. Rev.*, D23:347–356, 1981; A. H. Guth and P. J. Steinhardt, The inflationary universe, *Sci. Amer.*, 250(5):116–128, May 1984; L. Krauss, Cosmological antigravity: The case for a cosmological constant, *Sci. Amer.*, 282(1):52–59, January 1999; L. Krauss, *Quintessence: The Mystery of the Missing Mass in the Universe*, 2000; A. D. Linde, Eternally existing self-reproducing chaotic inflationary universe, *Phys. Lett.*, B175:395–400, 1986; A. D. Linde, The inflationary universe, *Rep. Prog. Phys.*, 47:925–986, 1984; D. N. Spergel et al., First-year *Wilkinson Microwave Anisotropy Probe* (*WMAP*) observations: Determination of cosmological parameters, *Astrophys. J. Suppl.*, 148:175–194, 2003; D. N. Spergel et al., *Wilkinson Microwave Anisotropy Probe* (*WMAP*) three year results: Implications for cosmology, *Astrophys. J.*, in press, 2006.

Inflorescence

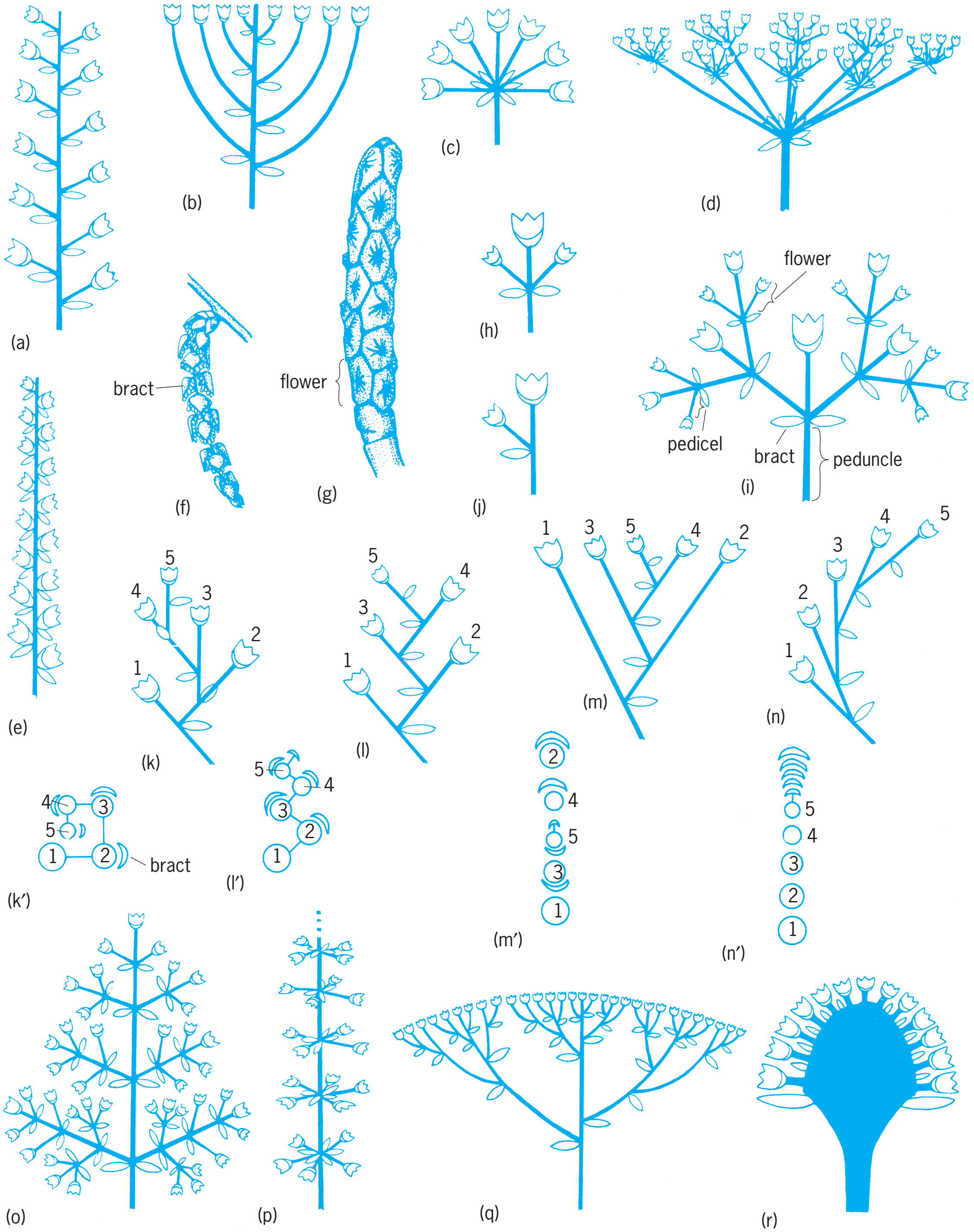
A flower cluster segregated from any other flowers on the same plant, together with the stems and bracts (reduced leaves) associated with it. Certain plants produce inflorescences, whereas others produce only solitary flowers. The stalks of the individual flowers are the pedicels.

Inflorescences are easily identified in certain species, though not in others. For example, it is often difficult, if not arbitrary, to determine whether flowers on a plant are sufficiently segregated from each other to be interpreted as solitary flowers or components of one or more inflorescences. In other cases, such as species of *Solidago* (goldenrod), a plant produces flowers in small clusters conventionally designated as inflorescences, but these inflorescences are aggregated into a very conspicuous inflorescencelike cluster at the tip of the plant.

Branching. Patterns of branching in inflorescences, as well as other portions of the plant, are classified under two general categories, monopodial and sympodial. Distinctions between these categories are not absolute, and a given inflorescence may exhibit one or both patterns. In monopodial branching, each branch is subordinate to that which bears it, and stems may bear numerous lateral branches (including flowers; **illus.** *a* and *o*). In sympodial branching, each branch is more or less equivalent to the branch which bears it, generally terminates in a flower, and tends to be short, bearing one or a few lateral branches (illus. *h*–*n*′). In some cases of sympodial branching, a structure which superficially appears as a single branch (that is, one derived from a single apical meristem) is actually composed of several branches.

Classification. Inflorescences are classified according to gross appearance at maturity or pattern of branching. Depending on the species, an inflorescence may be regarded as belonging to one or more types; for example, the inflorescence of *Limnocharis flava* (velvet leaf) is properly classified as an umbel (based on gross appearance), yet also as a compound monochasium or bastryx (based on branching pattern).

Types of inflorescences include: raceme (illus. *a*), composed of a monopodial central axis, along the sides of which arise pedicel-bearing flowers; corymb (illus. *b*), a raceme whose pedicels are proportionally elongated so as to bring all the flowers more or less to the same level; umbel (illus. *c*), composed of several flowers which radiate from almost the same point; compound umbel (illus. *d*), composed of several branches which radiate from almost the same point, each branch terminating in an umbellike flower cluster; spike (illus. *e*), composed of flowers which lack or virtually lack pedicels, arranged singly or in contracted clusters along a central axis; catkin (ament; illus. *f*), a slender, usually scaly spike bearing unisexual, apetalous flowers, which often hangs downward and is deciduous as a whole; spadix (illus. *g*), a spike with a thickened fleshy axis, usually enveloped by one or more spathes (large modified leaves); dichasium (simple dichasium; illus. *h*),



Examples of inflorescence types. All are side views except *k*′, *l*′, *m*′, and *n*′ which are top views. Numbers in *k*–*n*′ indicate flowers in order of development. (*a*) Raceme. (*b*) Corymb. (*c*) Umbel. (*d*) Compound umbel. (*e*) Spike. (*f*) Catkin. (*g*) Spadix (fruiting, *Ludovia integrifolia*). (*h*) Dichasium. (*i*) Compound dichasium. (*j*) Monochasium. (*k*, *k*′) Bostryx. (*l*, *l*′) Cincinnus. (*m*, *m*′) Rhipidium. (*n*, *n*′) Drepanium. (*o*) Panicle. (*p*) Verticillate inflorescence. (*q*) Cyme. (*r*) Head.

a three-flowered cluster composed of an axis which bears a terminal flower and, below it, two bracts which subtend lateral flowers; compound dichasium (illus. *i*), composed solely of dichasia; monochasium (simple monochasium; illus. *j*) a two-flowered cluster composed of an axis which bears a terminal flower and, below it, one bract which subtends a lateral flower; compound monochasium (illus. *k*–*n*′), composed solely of monochasia; bostryx (helicoid cyme; illus. *k* and *k*′), a compound monochasium or modified compound dichasium in which successive pedicels describe a spiral; cincinnus (scorpiod cyme, illus. *l* and *l*′), like a bostryx, but successive pedicels follow a zigzag path; rhipidium (illus. *m* and *m*′), a bostryx compressed so that all flowers lie in one plane; drepanium (illus. *n* and *n*′), a cincinnus compressed so that all flowers lie in one plane; panicle (illus. *o*), composed of a monopodial central axis which bears flowers indirectly on branches of higher order and which also may bear some flowers directly; thyrse, a compact panicle of more or less cylindrical form; verticillate inflorescence (illus. *p*), like either a raceme or panicle, but the central axis bears flowers in whorls, approximate whorls, or contracted clusters which simulate whorls; cyme (illus. *q*), a compound, more or less flat-topped sympodial inflorescence, sometimes simulating a compound umbel but having branches less regularly disposed; and head (capitulum; illus. *r*), compact inflorescence having a very short, often discoid or globose flower-bearing portion.

Formerly, the aforementioned inflorescence types were classified under two main categories, indeterminate (racemose) and determinate (cymose); however, this classification is now obsolete. *See* FLOWER.

George J. Wilder

Bibliography. H. C. Bold et al., *Morphology of Plants and Fungi*, 1980; H. W. Rickett, The classification of inflorescences, *Bot. Rev.*, 10:187–231, 1944; H. W. Rickett, Materials for a dictionary of botanical terms, III: Inflorescences, *Bull. Torrey Bot. Club*, 82:419–445, 1955.

Influenza

An acute respiratory viral infection characterized by fever, chills, sore throat, headache, body aches, and severe cough; it is commonly known as the flu. While many viruses can cause respiratory infections such as the common cold, influenza viruses are more likely to cause severe illness and to result in serious medical complications. Although gastrointestinal symptoms may sometimes accompany influenza infection, especially in children, the term "stomach flu" is a misnomer for gastrointestinal illnesses caused by a variety of other viruses, bacteria, or other agents. *See* VIRUS.

Another unique feature of influenza is its distinctive seasonal pattern. Outbreaks of influenza occur during cold weather months in temperate climates, and typically most cases cluster during a period of 1–2 months, in contrast to broader periods of circulation with many other respiratory viruses. During annual influenza epidemics, an increased death rate due to influenza-related complications is often observed, and is one of the ways that public health authorities monitor epidemics. During an average influenza season in the United States, more than 200,000 people are hospitalized and 36,000 die as a result of such complications. *See* PUBLIC HEALTH.

Viral agents. The bacterium *Hemophilus influenzae* was once mistakenly considered the cause of influenza, based on finding the microbe in the sputum of individuals with the disease. However, viruses were shown to be the cause in 1933, when the first influenza virus, designated type A, was isolated. The second major type, designated type B, was identified in 1940, and the third, type C, in 1949.

The three types of influenza viruses are classified in the virus family Orthomyxoviridae, and they are similar, but not identical, in structure and morphology. Types A and B are more similar in physical and biologic characteristics to each other than they are to type C; and unlike type C viruses, they are responsible for widespread illness during epidemics. Influenza viruses may be spherical or filamentous in shape, and they are of medium size among the common viruses of humans. There are two important viral protein antigens on the virion surface, hemagglutinin and neuraminidase. Hemagglutinin functions in the attachment of the virus to cell receptors and initiates infection by anchoring the virus to respiratory epithelial cells. Neuraminidase is an enzyme that facilitates release of virus from infected cells and aids in the spread of virus in the respiratory tract. *See* ANTIGEN.

Influenza type A viruses are divided into subtypes based on the hemagglutinin and neuraminidase proteins. There are 16 different hemagglutinin subtypes and 9 neuraminidase subtypes. Currently there are two influenza A subtypes circulating among humans, but all hemagglutinin and neuraminidase subtypes have been found in wild birds, which are believed to be the natural reservoir for all influenza A subtypes. Influenza type B does not have an animal reservoir and is not divided into subtypes.

Antigenic drift and shift. Influenza type A and B viruses undergo changes over time due to an accumulation of point mutations in the viral proteins, especially hemagglutinin and to a lesser extent neuraminidase. This is referred to as antigenic drift, and necessitates changes to at least one of the three virus strains in the influenza vaccine almost every year. These changes are also responsible for the fact that people remain susceptible to influenza infections during their lives, since antibodies produced in response to infection with a particular influenza strain may not recognize the mutated viruses and therefore may provide little or no protection against them as the strains continue to mutate. Influenza viruses that currently cause epidemics among humans are influenza types A(H1N1), A(H3N2), and B. *See* MUTATION.

Influenza A viruses may also undergo sudden changes known as antigenic shift, which results in a virus with a new hemagglutinin, neuraminidase, or both. When this occurs, a new virus subtype emerges to which most or all people have no immunity, resulting in a worldwide influenza epidemic, or pandemic. Influenza pandemics occurred in 1918–1919, 1957–1958, and 1968–1969. The 1918–19 pandemic (the Spanish influenza pandemic) was by far the most severe, and is estimated to have caused 20–40 million deaths worldwide with 675,000 in the United States alone. The reasons for the unusually severe illnesses and high death rates during this pandemic have long puzzled influenza researchers, and for many years little was known about the characteristics of the causative virus. However, in recent years a team of scientists has recovered influenza virus genetic material (RNA) from formalin-fixed autopsy samples taken from soldiers who died of influenza during the 1918–1919 pandemic and from frozen lung tissues of a person who died of influenza and was buried in the permafrost (perennially frozen ground) in Alaska in 1918. The genetic material recovered from these sources has been used to reconstruct the 1918 Spanish influenza pandemic strain. Studies of this reconstructed virus will greatly enhance the understanding of that pandemic virus, and perhaps will help scientists better detect viruses with pandemic potential and develop strategies to help mitigate the effects of pandemic influenza in the future. *See* EPIDEMIC.

Animal influenza. In addition to infecting humans, strains of influenza A infect many other species, including pigs, horses, seals, whales, and a wide variety of avian species. The severity of influenza infection in animals ranges from asymptomatic to fatal.

Animal influenza viruses are believed to play an important role in the emergence of human pandemic strains. One way that antigenic shift can occur is when a human and an animal influenza A virus strain infect the same host. When a cell is infected by two different type A viruses, various combinations of the original parental viruses may be assembled into the new progeny; thus, a progeny virus may be a mixture of gene segments from each parental virus and therefore may gain a new characteristic, for example, a new hemagglutinin. Hence a new virus is introduced into a population that has no immunity against it.

Since pigs are susceptible to infection with both human and avian influenza strains, this type of antigenic reassortment could occur when a pig is simultaneously infected with human and avian influenza strains. This can happen in settings where people, pigs, and a variety of avian species live in proximity, a circumstance common in many parts of the world. Since in recent years it has been observed that humans can also become infected with some strains of avian influenza viruses, it is also theoretically possible for reassortment to occur if a human is simultaneously infected with an avian and a human influenza virus.

The first human illnesses associated with an outbreak of a highly pathogenic A(H5N1) virus in birds were reported in 1997. This virus was widespread among poultry throughout live bird markets in Hong Kong, and led to 18 known human illnesses and 6 deaths. Further spread of this virus in Hong Kong was prevented by destroying 1.5 million birds. In 2003 the A(H5N1) virus was discovered in poultry in Vietnam and has since spread to other countries in Asia, Europe, and Africa, where it has been found in poultry and wild birds. By July 2006 a total of 231 cases in humans had been confirmed in 10 countries, of which 133 were fatal. Although most human cases of "bird flu" are believed to have been contracted by contact with infected birds, it is possible that the virus might change, either by reassortment or mutation of the viral genes, and become easily transmitted from human to human, causing a pandemic. This has prompted governments and health agencies to rush development of vaccines and to develop pandemic preparedness plans as a precaution, as well as to intensify surveillance of this and other avian influenza viruses with similar potential, and to destroy infected poultry when outbreaks are detected. *See* ANIMAL VIRUS.

Pathogenesis. When influenza virus enters the respiratory tract, usually by inhalation of aerosolized (suspended in air) viruses from an infected person, the virus hemagglutinin attaches to epithelial cells and invades the cells. After attachment, the cell's normal defense mechanisms actually help the virus gain entry and replicate. The cell engulfs the virus and attempts to destroy it with enzymes. However, instead of destroying the virus, the enzymes allow the viral RNA proteins to spill into the cell and move to the cell nucleus. Replication of viral RNA begins rapidly in the nucleus of the host cell. Final assembly of new virus occurs at the plasma membrane, and new viruses spread to nearby cells. The virus has a short incubation period. There is only a period of 1–3 days between infection and illness, and this leads to the abrupt development of symptoms that is a hallmark of influenza infections. The virus is typically shed in the throat for 5–7 days, during which time the sick person can infect others. Complete recovery from uncomplicated influenza usually takes several days to a week, and the individual may feel weak and exhausted for a week or more after the major symptoms disappear.

Numerous medical complications can occur as a result of influenza infection. Pulmonary complications include secondary bacterial and primary viral pneumonias. Viral pneumonia is far less common than bacterial pneumonia, but the mortality rate is much higher. However, because it occurs much more frequently, bacterial pneumonia is responsible for many more hospitalizations and deaths during influenza epidemics. Other pulmonary complications of influenza include croup in infants and young children and worsening of chronic lung conditions. Even previously healthy people can show alterations in pulmonary function that can last for weeks after the initial infection. Influenza can also cause inflammation of the heart muscle (myocarditis) or of the sac around the heart (pericarditis), and may worsen chronic heart conditions. It can also cause encephalitis (inflammation of the brain), kidney failure, and Reye's syndrome, which affects the liver and central nervous system. Reye's syndrome occurs almost exclusively in children, and has been associated with the use of aspirin to treat influenza (as well as chickenpox). For this reason it is recommended that aspirin substitutes such as acetaminophen be used when treating children for suspected viral infections. *See* ASPIRIN; HERPES; PNEUMONIA.

Prevention and control. Vaccination is the primary means of preventing influenza. Because at least one of the three virus strains in the influenza vaccine is changed during most years due to viral mutations, and because vaccine-induced antibody can wane over the course of a year, the vaccine must be taken every year before the influenza season. Currently there are two types of influenza vaccines. One is made from inactivated, or killed, viruses, and the other from living viruses that have been attenuated, or weakened. Anyone over 6 months of age can take the inactivated vaccine, but the live attenuated vaccine is recommended only for those between the ages of 5 and 49 years who are not pregnant. The inactivated vaccine is administered by injection, but the live attenuated vaccine is delivered as a nasal spray. With both vaccines, antibodies develop within about 2 weeks of vaccination, but the live virus vaccine may better stimulate the cell-mediated immune response and provide broader and longer-lasting immunity, especially in children and younger adults. How effective the vaccine is in preventing influenza also depends on how healthy the immune system of the recipient is, and on how closely the virus strains in the vaccine resemble those that circulate during the influenza season. Since the vaccine strains must be chosen well in advance of the influenza season, one or more viruses may continue to mutate to an extent that causes the vaccine to be less effective. However, even when this happens, people who have been vaccinated usually have a shorter and less severe illness if they are infected, and are less likely to develop complications than people who were not vaccinated.

Antiviral medications can be used to prevent or to treat influenza. These drugs act by inhibiting replication of the virus and, to prevent influenza, must be taken daily during the period of possible exposure. The drugs can also decrease the severity and lessen the duration of the illness if taken soon after symptoms begin. Two of these drugs, amantadine (Symmetrel) and rimantadine (Flumadine), classed as M2 inhibitors (since they inhibit viral protein M2, which uncoats the virus's protein shell, a process needed for viral replication in the host cell), are effective only against type A viruses. While the drugs were used effectively for many years, influenza viruses can develop resistance during treatment, and in recent years so many resistant strains have been detected that the two are no longer the drugs of choice. Newer drugs, Osletamivir (Tamiflu) and Zanamivir (Relenza), classified as neuraminidase inhibitors, are used to treat influenza with relatively few problems of resistance to date. These drugs are active against both type A and type B viruses. *See* BIOTECHNOLOGY; VACCINATION. John M. Quarles; Nancy H. Arden

Bibliography. J. A. Cabezas et al., New data on influenza virus type C confirm its peculiarities as a new genus, *Intervirology*, 32:325–326, 1991; R. B. Couch et al., Influenza: Its control in persons and populations, *J. Infect. Dis.*, 153:431–440, 1986; J. M. Katz, Preparing for the next influenza pandemic, *ASM News*, 70:412–419, 2004; E. D. Kilbourne, *Influenza*, 1987; M. W. Shaw, N. H. Arden, and H. F. Maassab, New aspects of influenza viruses, *Clin. Microbiol. Rev.*, 5:74–92, 1992; T. M. Tumpey et al., Characterization of the reconstructed 1918 Spanish influenza pandemic virus, *Science*, 30:77–80, 2005.

Information management

The functions associated with managing the information assets of an enterprise, typically a corporation or government organization. Increasingly, companies are taking the view that information is an asset of the enterprise in much the same way that a company's financial resources, capital equipment, and real estate are assets. Properly employed, assets create additional value with a measurable return on investment. Forward-looking companies carry this view a step further, considering information as a strategic asset that can be leveraged into a competitive advantage in the markets served by the company.

In some respects, the term "information management" is interchangeable with management information systems (MIS); the choice of terminology may reflect the attitude of a company regarding the value of information. Companies using the traditional MIS descriptor are likely to hold the view that management information systems represent a cost to be minimized, in contrast to the strategic-asset view with the objective of maximizing return on investments.

For the large enterprise with multiple business units and functional elements distributed over a broad geographic area, information management can be an enormously complex task. Its successful accomplishment is dependent not only on the diligent application of professional skills, but even more on leadership and management abilities strongly rooted in a thorough knowledge of and insight into the enterprise.

Scope. The scope of the information management function may vary between organizations. As a minimum, it will usually include the origination or acquisition of data, its storage in databases, its manipulation or processing to produce new (value-added) data and reports via application programs, and the transmission (communication) of the data or resulting reports. While many companies may include, with good reason, the management of voice communications (telephone systems, voice messaging, and, increasingly, computer-telephony integration or CTI), and even intellectual property and other knowledge assets, this article focuses primarily on the data aspects of information management.

Information versus data. Executives and managers frequently complain of "drowning in a flood of data, but starving for information." There is a significant difference between these two terms. Superficially, information results from the processing of raw data. However, the real issue is getting the right information to the right person at the right time and in a usable form. In this sense, information may be a perishable commodity. Thus, perhaps the most critical issue facing information managers is requirements definition, or aligning the focus of the information systems with the mission of the enterprise. The best technical solution is of little value if the final product fails to meet the needs of users.

Information engineering. One formal approach to determining requirements is information engineering. By using processes identified variously as business systems planning or information systems planning, information engineering focuses initially on how the organization does its business, identifying the lines from where information originates to where it is needed, all within the context of a model of the organization and its functions. While information systems personnel may be the primary agents in the information engineering process, success is critically dependent on the active participation of the end users, from the chief executive officer down through the functional staffs.

An enterprise data dictionary is a principal product of the information engineering effort. This dictionary is a complete catalog of the names of the data elements and processes used within the enterprise along with a description of their structure and definition. The dictionary may be part of a more comprehensive encyclopedia, which is typically a formal repository of detailed information about the organization, including the process and data models, along with descriptions of the design and implementation of the organization's information systems applications.

A major advantage of the application of information engineering is that it virtually forces the organization to address the entire spectrum of its information systems requirements, resulting in a functionally integrated set of enterprise systems. In contrast, ad hoc requirements may result in a fragmented set of systems (islands of automation), which at their worst may be incompatible, contain duplicate (perhaps inconsistent) information, and omit critical elements of information.

Information systems architecture. Information engineering develops a logical architecture for an organization's information systems. Implementation, however, will be within the context of a physical systems architecture. Information engineering ensures functional integration, whereas systems engineering provides for the technical integration of information systems. Together, functional and technical (logical and physical) integration lay the architectural foundation for the integrated enterprise. *See* SYSTEMS ARCHITECTURE; SYSTEMS ENGINEERING.

Open versus proprietary architectures. Many critical decisions are involved in establishing a physical systems architecture for an organization. Perhaps the single most important decision is whether the architecture will be based on a single manufacturer's proprietary framework or on an open systems framework. Although the former may be simpler to implement and may offer less technical risk, it is far more restrictive with regard to choices of equipment. An open systems architecture allows for the integration of products from multiple vendors, with resulting increased flexibility in functional performance and the potential for significant cost savings in competitive bidding. The flexibility to incorporate new products is particularly important in an era of rapid technological advancement, where new technology may become obsolete in 18 months.

Client-server and network computing. One of the most widely used architectural models is the client-server model. In this model, end users, for example, may employ personal computers as clients to request data from an enterprise database resident at another location on a minicomputer or mainframe (the server). Clients and servers interconnect and communicate with one another over networks, from local-area networks (LANs) within a single facility, to wide-area networks (WANs) which may span one or more continents. A variation on this model (sometimes referred to as network computing) employs "thin" clients, with the applications running on servers accessed over the network. *See* CLIENT-SERVER SYSTEM; DISTRIBUTED SYSTEMS (COMPUTERS); LOCAL-AREA NETWORKS; WIDE-AREA NETWORKS.

Operating systems. An extremely important architectural consideration is the choice of operating systems. This selection may determine the options available in acquiring off-the-shelf software packages. Alternatively, an existing (legacy) set of software application programs (normally representing a major investment) may severely constrain the choice of hardware on which the programs will run. The issue here is whether the operating systems are proprietary to a single vendor's hardware or are vendor-independent (such as UNIX). Because of the portability it offers, UNIX (and more recently a derivative of UNIX called LINUX) has become a de facto standard in the industry. Although it is actually proprietary, Microsoft's Windows family of operating systems has also achieved de facto standard status due to its increasingly ubiquitous availability on most major hardware platforms. *See* OPERATING SYSTEM.

Databases. Another major architectural decision is the structure of the databases to be utilized. Most database products follow a hierarchical, relational, or object-oriented paradigm. Although the most widely used structure has been relational, increasing attention is being given to object-oriented systems. Information engineering tends to separate data from applications, while object-oriented methods couple the two. *See* DATABASE MANAGEMENT SYSTEM; OBJECT-ORIENTED PROGRAMMING.

Internet and related technologies. The rapid rise of the Internet since the mid-1990s has profoundly impacted information systems architectures as well as the role of information systems within an organization. With tens of millions of individuals, corporations, and even governments tied in to the Internet, there are few businesses that do not reach out to their customer base via a Web site, thus expanding their information systems architecture to encompass the extended enterprise. Equally important, the same technology that drives the Internet (the Internet Protocol or IP) has made possible intranets, which function like the Internet but are private to the organization. The corporate Web site involves the marketing and public relations organizations as major information systems users, and even the sales organization may be involved, as actual sales are made over the Web, giving rise to "electronic commerce" (e-commerce). *See* INTERNET; WORLD WIDE WEB.

Software applications. The information engineering process can produce a complete set of requirements for the applications needed to run the enterprise. When all architectural parameters have been determined, unique applications can be developed, preferably employing computer-aided software engineering (CASE) tools to automate the process to the maximum degree possible. In general, however, the cost and time involved in developing custom software, and the even greater cost of supporting (maintaining) that software throughout its life cycle, dispose strongly toward the maximum use of off-the-shelf software packages. *See* SOFTWARE ENGINEERING.

There are many functions that are common to almost any organization, such as personnel (human resources) and finance. Applications that support these and other basic infrastructure functions, such as manufacturing, are readily available off-the-shelf from major vendors in the form of enterprise resource planning (ERP) software packages that have become a major element of the software industry. Although considerable work is required either to configure the vendor's software to match the organization's internal processes, or to alter the processes to match the software, or a combination of the two, the net cost in time and dollars (including maintenance) often makes this option more attractive than developing custom applications from scratch.

Functions that are unique to an organization may justify the cost and risk associated with custom software development, but these decisions should not be made without careful analysis. Strategic applications that produce a compelling competitive advantage are most likely to meet this test.

Enterprise application integration. The proliferation of Web and e-commerce applications, as well as the trend toward enterprise resource planning systems, has posed an additional major challenge for information managers: successfully integrating these new applications with the enterprise's legacy systems, or enterprise application integration (EAI). *See* INFORMATION SYSTEMS ENGINEERING; SYSTEMS INTEGRATION.

Alan B. Salisbury

Bibliography. S. Alter, *Information Systems: A Management Perspective*, 1999; J. Martin, *Information Engineering*, 1989; A. Sage, *Systems Management for Information Technology and Software Engineering*, 1996; T. Wheeler, *Open Systems Handbook*, 1992.

Information processing (psychology)

The coding, retrieval, and combination of information in perceptual recognition, learning, remembering, thinking, problem solving, and performance of sensory-motor acts.

Coding. Information is coded in some type of memory by having something in the memory represent the information. In books, recording tapes, and almost all digital computer memories, the memory consists of a long one-dimensional series of memory locations. The encoding of a piece of information such as the proposition "John plays tennis" consists of imposing the patterns representing "John," "plays," and "tennis" into adjacent locations in the memory. Any other piece of information, such as the proposition "Tennis is a sport," learned at a different time would be coded in a similar manner in a totally different set of locations in memory. The same pattern would be used to represent the concept "tennis," which is common to both propositions, but this pattern would be stored in two separate places, and an information processor using the memory would not have direct access to both facts about "tennis" when it had retrieved one fact. An additional search of memory would be required. Such a memory is called nonassociative, because there are no direct associations linking all instances of the same concept or idea in memory.

Human memory is not at all like a tape recorder or computer memory. Human memory is associative. From a coding standpoint, the defining property of an associative memory is the use of specific node encoding, in which each idea is represented only once by a designated node (location). The representative of an idea is a particular location in memory, rather than a particular pattern that can be, and often is, stored in a variety of locations. There are many different types of associative memories; two of these are considered below: horizontal associative memory and vertical associative memory.

In a horizontal associative memory, the coding of propositions consists of forming a strong association directly from each concept node to the following concept node (**Fig. 1**). From a semantic viewpoint, the association from one node to another means "is followed by." Careful study of Fig. 1 reveals the two primary deficiencies of a horizontal associative memory. First, there is no encoding of where one proposition ends and another begins, a difficulty which might be referred to as "run-on propositions." Second, there is a serious associative interference problem. For example, there is no way to know whether John plays tennis, basketball, or both, and similarly for Peter. The ability of an associative memory to integrate information that involves the same concepts can produce excessive and erroneous integration in the case of a horizontal associative memory.

Vertical associative memories solve these problems while retaining the desirable feature of specific node encoding. They associate atomic concepts to higher-order chunk (phrase) concepts. These associations mean that the lower-order concept "is a constituent of" the higher-order (phrase or propositional) concept. All of the concepts in a single proposition are associated directly or indirectly to a single higher-order node that represents the entire proposition (**Fig. 2**). In this way both the run-on proposition and associative interference problems are solved by a hierarchical tree-structure encoding of exactly which concepts belong in each proposition.

Associations are assumed to be unidirectional, that is, from one node to another. Information processing in an associative memory requires the ability to get from a chunk node to its constituents, and vice versa. Semantically, the associations from a chunk to its constituents mean "has as a constituent." Since the meaning of vertical associations in the two directions is logically opposite, it seems likely that these associations are represented in the mind (brain) by structurally dissimilar types of links that function somewhat differently when information is retrieved from a person's memory. Many theories of associative memory assume an even larger variety of links than this, but at least these two types appear to be necessary.

Order and context-sensitive coding. There are many chunks whose constituents have no unique ordering, such as the members of a family or the features of many objects. Although the sentences used to communicate propositions to others have a definite word order, the same proposition can be communicated using virtually any ordering of the constituent concepts, provided the proper syntactic plan is used. Semantically, there does not appear to be any unique ordering of the constituents of a proposition. Hence, it is likely that the encoding of propositions is another example of the encoding of an unordered set.

However, there are numerous examples of chunks encoded in the mind whose constituents are uniquely ordered, for example, the episodes of a story or the letters in a word. How are these ordered sets coded? A number of possibilities have been suggested, but the one that appears to have the most evidence supporting it is context-sensitive coding. According to this theory, the constituents of a chunk representing an ordered set such as the word "clank" are nodes representing overlapping triples of the component segments. In the case of "clank" the constituents would be "#Cl," "cLa," "lAn," "aNk," and "nK#." The reason for distinguishing the middle element in each triple is that it is the central or focal element and the two on each side are the context. In speech production, the node "lAn" would control the speech musculature to produce the particular "a" that is preceded by "l" and followed by "n." If horizontal associations are added to the constituents of a chunk as a third type of link in vertical associative memory, then horizontal associations between pairs of context-sensitive nodes will produce a unique ordering of constituents when there is one and no such ordering when there is not. In the "clank" example, sufficient information has been encoded to know the first constituent of the set, given that the entire set is active in the mind due either to bottom-up activation by stimulus presentation of the word "clank" or to top-down activation of the segmental constituents by the word node for "clank." In both recognition and production of the work "clank," there is only one segmental node of the form "#Xy." Since "#" represents the junction between words, this must be the first node in the word. Given that "#Cl" is the first node, "cLa" will have the strongest horizontal association from "#Cl," so one knows it is the second node, and so forth.

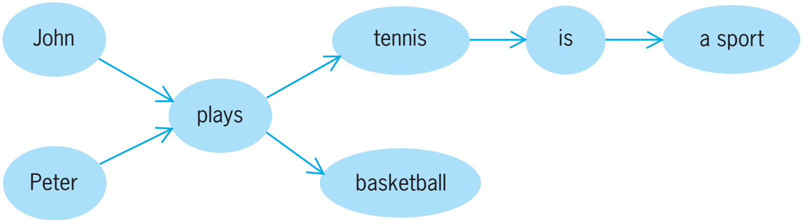


Fig. 1. Horizontal associative memory with a serious associative interference problem (who plays tennis—John, Peter, or both?) and no encoding of where each proposition begins and ends.

Learning. All learning in an associative memory can be considered to be the strengthening of different associations between nodes. Thinking of a set of ideas at about the same time (within approximately a second) activates the nodes representing the ideas at about the same time. In a way that is not understood, this strengthens vertical associations gluing these idea nodes to higher-order chunk nodes and probably also strengthens horizontal associations among the idea nodes. If a set of idea nodes is simultaneously activated, then all horizontal associations are equally strong on the average (all orders are encoded), and no unique ordering is encoded. If the set of idea nodes is activated in a simple linear order, than a unique set of horizontal associations will be encoded to represent the sequence (ordered set).

For certain purposes it may be useful to distinguish among three types of learning: chunking (forming

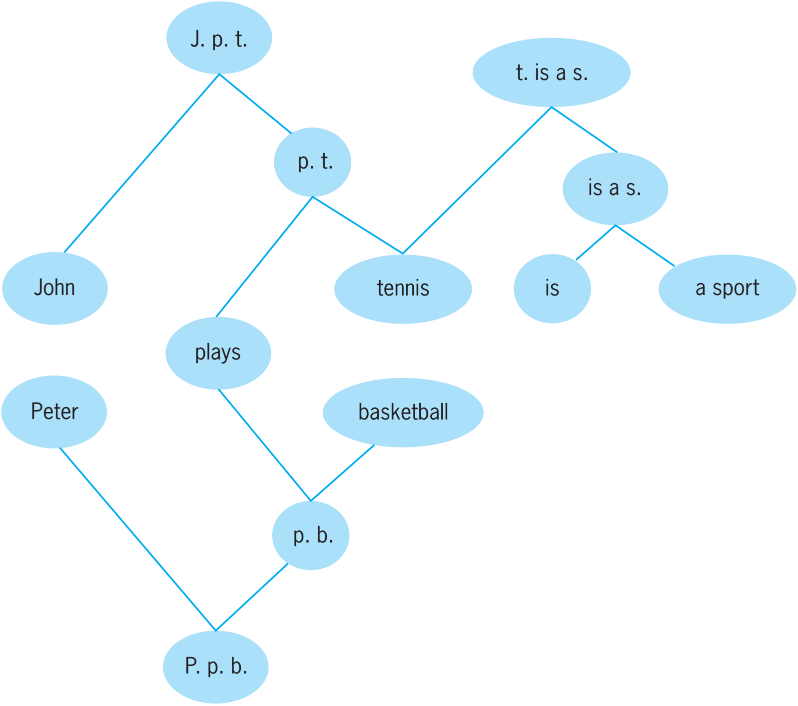


Fig. 2. Vertical associative memory where concepts are grouped into phrases and propositions by associating each concept to a higher-order node that represents the entire group.

new vertical associations from a set of old nodes to a new node, specifying that node to represent the set of old nodes); sequencing (forming new horizontal associations between adjacent pairs of nodes in an ordered set); and strengthing (increasing the strength of existing associations). Forming new associations refers to the initial increase in strength above the baseline, or zero, value, so all learning consists of the strengthening of associations in the most general sense.

However, the initial strengthening of associations appears to require more thinking than does the further strengthening of previously learned associations. Chunking and sequencing are greatly facilitated by coding new material so as to integrate it with stored material as much as seems relevant and as consistently as possible. When such integration uses relationships that are true and of considerable general value, it is called understanding. When such integration uses fictional relationships or relationships of little general value, it is called a mnemonic trick. Either way, new learning is made easier by integration with prior knowledge. The difference is that all associations formed or strengthened by understanding are true and useful, as opposed to only a few of the associations formed or strengthened by mnemonic tricks.

Further strengthening of learned associations requires much less thinking effort—just retrieval of the associations to be rehearsed. However, this does not mean that such review is of no value. Review can serve to counteract the effects of prior forgetting, to overlearn so as to retard future forgetting, to increase retrieval speed for the reviewed material in the future, and to reduce the attentional demands of such retrieval.

Retrieval. Information that is currently being thought about is said to be in active memory. All of the rest of the information that is stored in long-term memory is said to be in passive memory. Retrieval refers to the cognitive processes that allow one to make use of stored information by converting the memory traces for particular pieces of information from the passive to the active state. Retrieval can be analyzed into two components: recognition, which is the activation of the nodes in memory that encode the cue information. Every act of retrieval requires recognition of the retrieval cues, and probably every act of retrieval involves recall of some information beyond that given in the retrieval cues.

The basic information-retrieval process in an associative memory is direct-access, that is, direct activation of the nodes encoding the cue and closely associated information. No search is required when the retrieval cues are sufficient to identify a unique memory node (location). When the retrieval cues are inadequate, there is sometimes information stored in long-term associative memory that would be adequate to retrieve a unique answer to the question. In such cases, a sequence of direct-access retrieval processes is initiated. For example, sometimes the associations from an acquaintance's face to his or her name is too weak to retrieve the name from the face cues. However, when the backward associations from the name to the face cues are added in, the information stored in memory is sufficient to recognize which name goes with that face. So one can engage in a memory search process that consists of recalling various names and judging the strength of association of each name with the face. The entire sequence is a search process, but each elementary act of retrieval in this complex sequence is a direct-access process. The whole purpose of having an associative memory is to make direct-access retrieval possible, by encoding every idea in one specific location that can be identified from the cue input without having to search through sets of alternative locations.

Retrieval dynamics. While retrieval from memory is direct, it is not instantaneous. One elementary act of retrieval requires approximately 1 s of time to reach completion, with a range from 0.3 to 3 s. Retrieval can be analyzed into three temporal phases: latent, transition, and terminal (**Fig. 3**). The latent phase extends from the time the retrieval cues are first presented until the initial time (*I*) that any response can be made which is at all influenced by the memory traces activated by the retrieval cues. The terminal phase is the period of time after retrieval is complete during which the retrieved set of memory traces is fully activated and able to control behavior and thought. In the terminal phase, retrieved memory strength is equal to the limit set by the passive strength level stored in long-term memory (*L*). The transition phase is the period in between, when retrieved memory strength increases from zero to *L*. This transition phase is not abrupt, as if retrieval occurred all at once. Instead, memory retrieval appears to occur incrementally over a period of many tenths of a second. If people respond to retrieval cues after some retrieval has occurred, but before the retrieval of the memory trace is complete, they respond at a level of accuracy which is greater than chance but less than they could achieve had they delayed their response until retrieval was complete. A plot

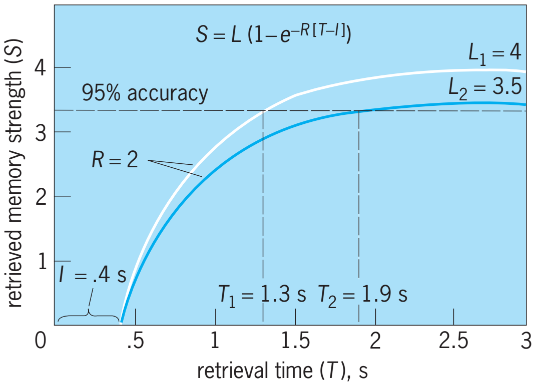


Fig. 3. Memory-retrieval functions for conversion of stored long-term memory strength *L* into active short-term memory strength *S* as a function of retrieval time *T*. Note that the time to achieve 95% accuracy is greatly affected by stored strength *L*, even when retrieval dynamics parameters (*I* + *R*) are identical.

of accuracy as a function of reaction time is called a speed-accuracy tradeoff function. The accuracy of a recognition or recall response to retrieval cues measures the strength of the retrieved memory trace. The increase in the strength of the retrieved memory trace as a function of the time allowed for processing the retrieval cues (retrieval time) is called the retrieval function. Empirical speed-accuracy tradeoff functions for recognition or recall thus provide measures of the theoretical retrieval functions for these tasks.

The dynamics of memory retrieval appears to be closely approximated by an exponential approach to a limit L at rate R after the initial delay time I (Fig. 3). Exponential approach to a limit means that in every very small increment of time (after delay time I), a constant percentage of what stored trace remains to be retrieved is retrieved. That percentage is the retrieval rate R. In Fig. 3, two retrieval functions are displayed that have identical retrieval dynamics (identical I and R parameters) but different limits set by their different levels of stored strength in long-term memory (L_1 and L_2). Note that the time required for memory retrieval to reach a certain level of accuracy (95% in Fig. 3) is considerably less for the trace with the greater strength in storage ($T_1 < T_2$). Thus, even without altering the dynamics of memory retrieval, one can decrease reaction time considerably in recall or recognition simply by increasing the strength of the stored trace in long-term memory, with no loss in accuracy.

Retrieval dynamics (I and R parameters) is not affected by memory strength. It is primarily affected by the hierarchical level of encoding of the memory trace being retrieved—the higher the coding level (that is, the longer the chain of upward vertical associations from the sensory periphery to the highest level chunk node), the slower the retrieval dynamics (that is, the greater the I value and the smaller the R value). Thus, the retrieval dynamics for recognizing letters is faster than that for words, which is in turn faster than that for sentences. The more simply one codes material in terms of hierarchical depth, the faster it can be retrieved, provided the strength of the trace is the same in both cases.

Attention and parallel processing. There are limits to how many things a person can attend to at once. The laws of attention are not completely understood, but the following appear to be true: A very large number of retrieval cues can be processed at once, provided they all converge on a single stored memory trace, regardless of complexity, which has been previously unified by vertical associations to a single chunk node at the top of its trace hierarchy; this is called convergent parallel processing. Two or three separate sets of retrieval cues can probably be processed, provided that the traces they retrieve are very strong; this is called divergent parallel processing. Divergent parallel processing is easier if the traces are stored in different memory modalities, for example, verbal versus spatial. Thus, a person can carry on a conversation and sign letters simultaneously, with little loss of effectiveness in either task, because the memory traces are in somewhat different modalities and the signature trace is very strong. One of the important aspects of increasing skill in sports, for instance, is the increasing ability to perform several information-processing tasks in parallel. Part of this skill is due simply to increased practice at each component skill and part is due to explicit coordination of these components under unified, higher-order chunk nodes. *See* COGNITION; MEMORY. Wayne A. Wickelgren

Bibliography. J. R. Anderson, *Cognitive Psychology and Its Implications*, 4th ed., 1999; K. Haberlandt, *Cognitive Psychology*, 2d ed., 1996; D. E. Rumelhart and J. L. McClelland, *Parallel Distributed Processing*, 1987; A. Trehub, *The Cognitive Brain*, 1991.

Information systems engineering

The process by which information systems are designed, developed, tested, and maintained. The technical origins of information systems engineering can be traced to conventional information systems design and development, and the field of systems engineering. Information systems engineering is by nature structured, iterative, multidisciplinary, and applied. It involves structured requirement analyses, functional modeling, prototyping, software engineering, and system testing, documentation, and maintenance.

Information systems. Modern information systems solve a variety of data, information, and knowledge-based problems. In the past, most information systems were exclusively data-oriented; their primary purpose was to store, retrieve, manipulate, and display data. Application domains included inventory control, banking, personnel record keeping, and the like. The airline reservation system represents the quintessential information system of the 1970s. Since then, expectations as to the capabilities of information systems have risen considerably. Information systems routinely provide analytical support to users. Some of these systems help allocate resources, evaluate personnel, and plan and simulate large events and processes. This extremely important distinction defines the range of application of modern information systems. The users expect information systems to perform all the tasks along the continuum shown in **Fig. 1**. In the 1980s the applications expanded to data-oriented and analytical computing.

Systems engineering. Systems engineering is a field of inquiry unto itself. There are principles of applied systems engineering, and a growing literature defines a field that combines systems analysis, engineering, and economics. Systems engineering extends over the entire life cycle of systems, including requirement definitions, functional designs, development, testing, and evaluation. The systems engineer's perspective is different from that of the product engineer, software designer, or technology developer. The product engineer deals with detail, whereas the systems engineer takes an overall viewpoint. Where the product engineer deals with internal operations,

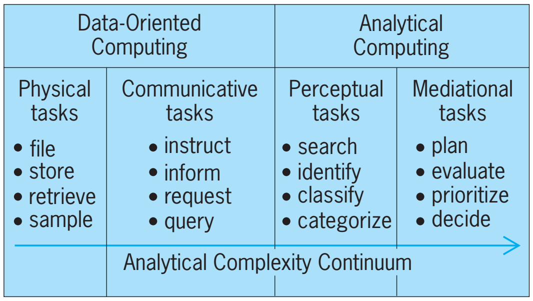


Fig. 1. Data-oriented and analytical computing, suggesting the range of information systems applications.

the systems engineer deals more extensively with the external viewpoint, including relations to other systems, users, repairers, and managers. Systems engineering is based upon the traditional skills of the engineer combined with additional skills derived from applied mathematics, psychology, management, and other disciplines. The systems engineering process is a logical sequence of activities and decisions that transform operational needs into a description of system performance configuration. *See* SYSTEMS ENGINEERING.

Figure 2 is a flowsheet for the design, development, and testing of information systems. The process is by nature iterative and multidisciplinary. Modern information systems problems are frequently analytical rather than data-oriented, and therefore require an iterative design, as the feedback loops in Fig. 2 suggest.

Requirements modeling and prototyping. The first and most important step is the identification of user requirements, on which software requirements are based; the two are by no means the same. Requirements analysis may be based on questionnaires and surveys, interviews and observation, and simulations and games.

Questionnaire and survey-based requirements analysis includes the use of importance-rating questionnaires, time-estimate questionnaires, and Delphi and policy capture techniques. Questionnaire and survey methods assume that insight into user requirements can be obtained from indirect queries.

Interview and field observation requirements analysis assumes that requirements can be determined by asking users what they do and how, and by observing them at the tasks that the system is to perform. These methods include the use of structured and unstructured interviews, the ad hoc convening of working groups, the use of critical incident techniques, and formal job analysis.

Simulation and gaming methods include the use of scenario-driven simulations of the functions of the system. Some simulations are paper-based, others are computer-based, and still others combine human and computer-based aspects of a problem. *See* SIMULATION.

Requirements analysis extends to the users of the system and the organization it is intended to support. **Figure 3** presents a three-dimensional requirements matrix comprising tasks, users, and organizational characteristics needed by information systems engineers.

Identified requirements should be modeled for validation; Fig. 2 suggests several methods. Popular techniques offer users a working demonstration of the fully programmed system. This is referred to as prototyping and may result in evolutionary and temporary systems. When requirements can be specified with some confidence, it is often possible to develop an evolutionary prototype that can be improved with time. When requirements are poorly defined, prototypes may be discarded repeatedly until requirements are clarified. *See* MODEL THEORY; PROTOTYPE.

System sizing. Once requirements have been modeled and validated, the information system concept can be specified. It requires formulation of a user–computer relation that will determine the user–system interaction, the specification of databases, the selection of analytical methods to drive data-oriented or analytical tasks, the identification of pertinent engineering principles, and the specification of hardware. *See* COMPUTER.

An important decision is the selection of a programming language. This may be determined by existing investments, experience of personnel, and limitations of hardware. Sometimes, however, the choice of language can be based upon requirements identified early in the process. Comparative programming languages is an emerging field that helps to make the choice. The growing range of high-level languages includes FORTRAN, C, Basic, COBOL, LISP, Prolog, Pascal, and Ada. *See* PROGRAMMING LANGUAGES.

Software engineering. Several techniques are available for the specification of software requirements and programming. There are data-flow, data-structure, and object-oriented tools and techniques, and a variety of methods for the generation of knowledge structures, modules, interfaces, and control structures. *See* DATAFLOW SYSTEMS; SOFTWARE ENGINEERING.

Testing, documentation, and maintenance. The information systems engineering process calls for testing throughout design and development. Of special importance is the testing of the software to determine if it satisfies user requirements, is consistent with specifications, and works efficiently. Quality assurance, fault tolerance, and redundancy techniques are available. *See* FAULT-TOLERANT SYSTEMS.

Information systems documentation consists of system specifications, functional descriptions, user manuals, and training materials.

System testing determines if the pieces of the system work harmoniously together; multiattribute utility and traditional cost–benefit methods are available. Systems are tested to determine whether they satisfy user requirements, are cost-effective, and can be maintained.

Information systems engineers also develop maintenance plans comprising schedules, procedures, and personnel.

Implementation. The introduction of a new information system into a computer-free environment can

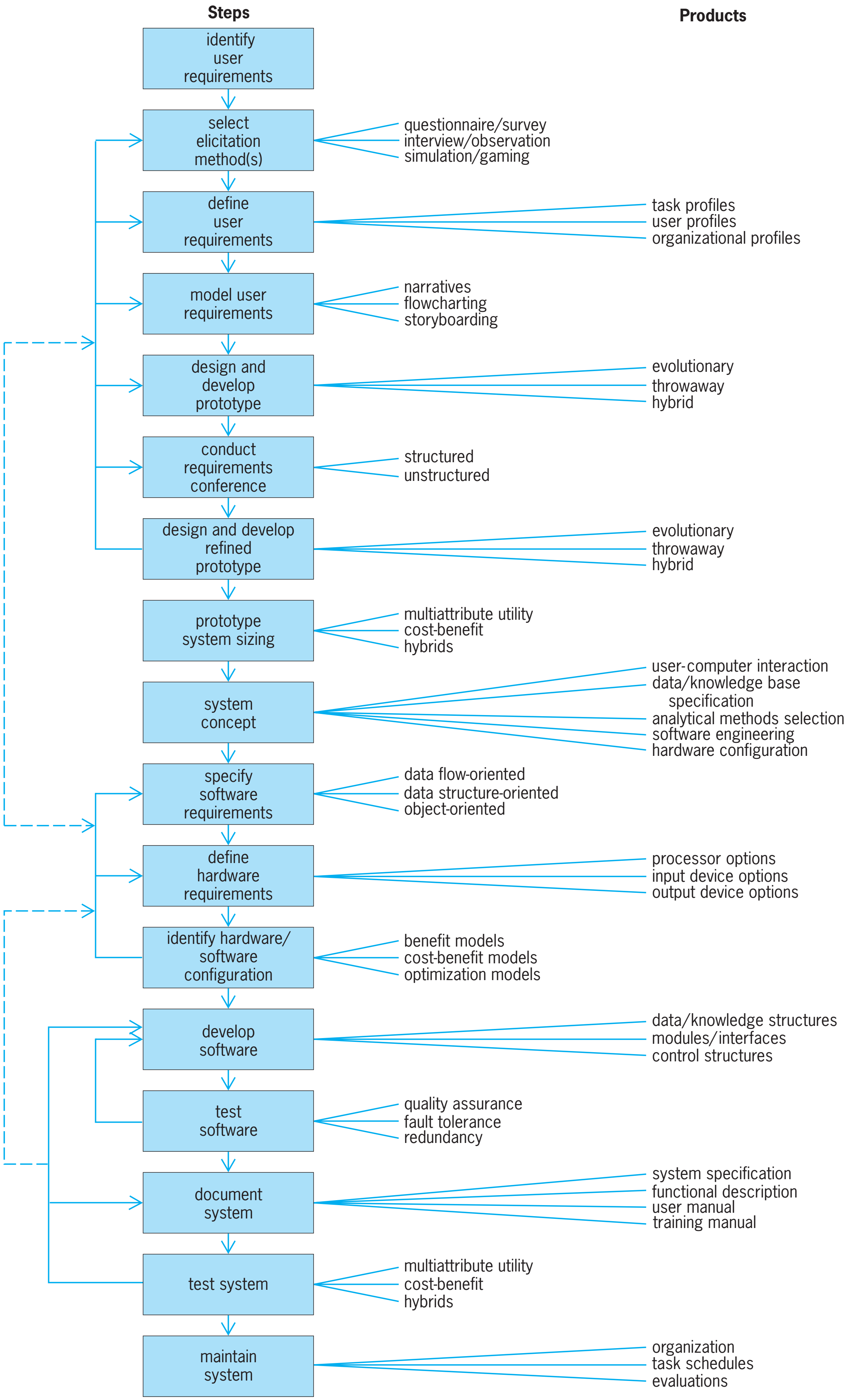


Fig. 2. Information systems engineering process, with steps, methods, and products.

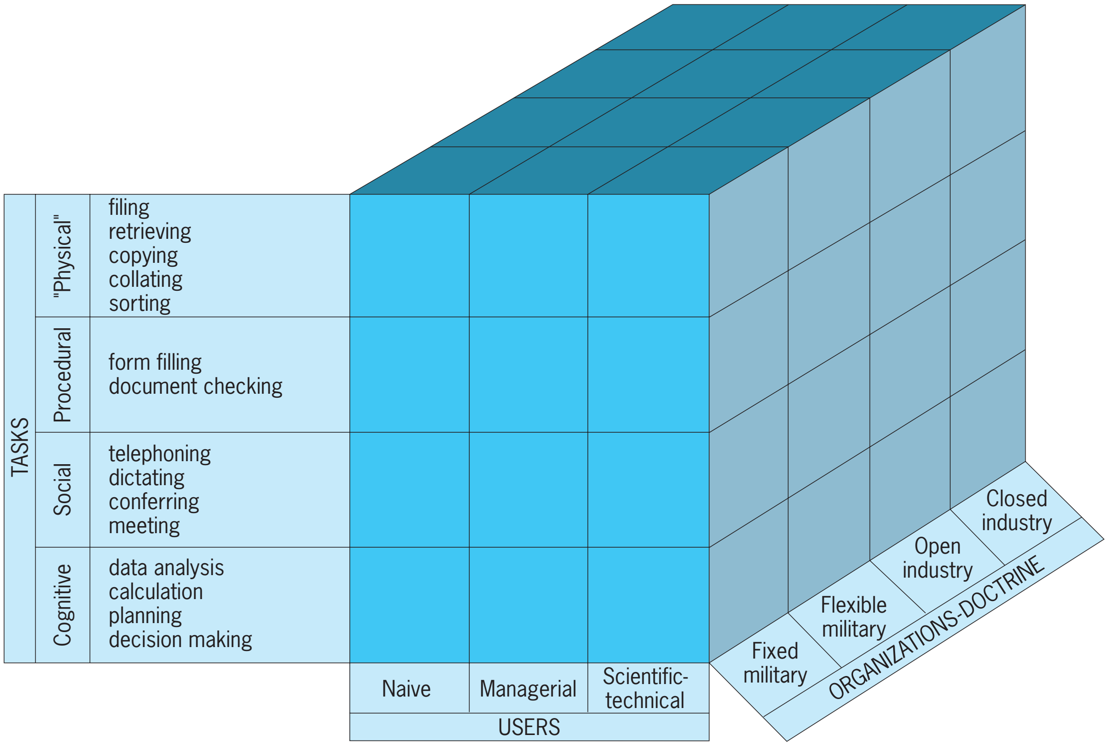


Fig. 3. Three-dimensional requirements matrix comprising information about tasks, users, and the organization that the system is intended to support.

create personnel problems. The information systems engineer adheres to procedures for the management of change developed by organizational theorists and sociologists.

Support. Information systems must be supported by an infrastructure group. This group is responsible for defining a maintenance and support process and providing the necessary tools to keep the application running and to plan for its modification, and, eventually, its replacement, triggering a whole new instance of the information systems engineering process.

Stephen J. Andriole

Bibliography. S. J. Andriole, *Managing Systems Requirements: Methods, Tools, and Cases*, McGraw-Hill, 1996; S. J. Andriole, *Storyboard Prototyping for Systems Design*: *A New Approach to User Requirements Validation and System Sizing*, 1989; R. S. Pressman, *Software Engineering*: *A Practitioner's Approach*, McGraw-Hill, 5th ed., 2000; A. P. Sage, *Systems Engineering*, Wiley-Interscience, 1992; A. Solvberg and D. C. Kung, *Information Systems Engineering: An Introduction*, Springer Verlag, 1993; K. M. Van Hee, *Information Systems Engineering: A Formal Approach*, Cambridge University Press, 1994.

Information technology

The field of engineering involving computer-based hardware and software systems, and communication systems, to enable the acquisition, representation, storage, transmission, and use of information. Successful implementation of information technology (IT) is dependent upon being able to cope with the overall architecture of systems, their interfaces with humans and organizations, and their relationships with external environments. It is also critically dependent on the ability to successfully convert information into knowledge.

Information technology is concerned with improvements in a variety of human and organizational problem-solving endeavors through the design, development, and use of technologically based systems and processes that enhance the efficiency and effectiveness of information in a variety of strategic, tactical, and operational situations. Ideally, this is accomplished through critical attention to the information needs of humans in problem-solving tasks and in the provision of technological aids, including electronic communication and computer-based systems of hardware and software and associated processes. Information technology complements and enhances traditional engineering through emphasis on the information basis for engineering.

Tools. Information technology at first was concerned with the implementation and use of new technologies to support office functions. These technologies have evolved from electric typewriters and electronic accounting systems to include very advanced technological hardware to perform functions such as electronic file processing, accounting, and word processing. Among the many potentially critical information technology–based tools are databases, e-mail, artificial intelligence systems, facsimile transmission (FAX) devices, fourth-generation programming languages, local-area networks (LANs), integrated service digital networks (ISDNs), optical disk storage (CD-ROM) devices, personal computers, parallel processing algorithms, word processing software, computer-aided software engineering packages, and accounting software.

Management information systems and decision support. The development of support through management information systems (MIS) became possible through the use of mainframe computers. These systems have become quite powerful and are used for a variety of purposes, such as scheduling airplane flights and booking passenger seats, and registering university students in classes. As management information systems began to proliferate, it soon was recognized that, while they were very capable of providing support for organizing data and information, they did not necessarily provide much support for tasks involving human judgment and choice. These tasks range from providing support in assessing situations, such as better detection of issues or faults, to supporting diagnosis in order to enable the identification of likely causative or influencing factors. This capability was provided by linking the database management systems (DBMS) so common in the MIS era with a model base management system (MBMS) capability and a visualization and interactive presentation capability made possible through dialog generation and management systems (DGMS). The resulting systems are generally known as decision support systems (DSS). Continuing developments in microchip technology have led to electronic communications–based networking, a major facet of information technology. Many would argue that the major information technology development in recent times is the Internet. *See* DECISION SUPPORT SYSTEM; INFORMATION SYSTEMS ENGINEERING.

Activities. The knowledge and skills required in information technology come from the applied engineering sciences, especially information, computer, and systems engineering sciences, and from professional practice. Professional activities in information technology and in the acquisition of information technology systems range from requirements definition or specification, to conceptual and functional design and development of communication and computer-based systems for information support. They are concerned with such topics as architectural definition and evaluation. These activities include integration of new systems into functionally operational existing systems and maintenance of the result as user needs change over time. This human interaction with systems and processes, and the associated information processing activities, may take several diverse forms. *See* REENGINEERING; SYSTEMS ARCHITECTURE; SYSTEMS ENGINEERING.

The hardware and software of computing and communications form the basic tools for information technology. These are implemented as information technology systems through use of systems engineering processes. While information technology and information systems engineering does indeed enable better designs of systems and existing organizations, it also enables the design of fundamentally new organizations and systems such as virtual corporations. Thus, efforts in this area include not only interactivity in working with clients to satisfy present needs but also awareness of future technological, organizational, and human concerns so as to support transition over time to new information technology–based services.

Systems integration. Often, it is very difficult to cope with the plethora of new information technology–based support systems. The major reason is the lack of systems integration across the large variety of such products and services. This has led to the identification of an additional role for information technology professionals, one involving support through information systems integration engineering. An information systems integration engineer is responsible for overall systems management, including configuration management, to ensure that diverse products and services are identified and assembled into total and integrated solutions to information systems issues of large scale and scope. There are many contemporary technological issues here. There is a need for open systems architectures, or open systems environments, that provide, for example, for interoperability of applications software across a variety of heterogeneous hardware and software platforms. The key idea here is the notion of open, or public, that is, intended to produce consensus-based developments that will ameliorate difficulties associated with lack of standards and the presence of proprietary interfaces, services, and protocols. *See* SYSTEMS INTEGRATION.

Distributed collaboration. In this network age of information and knowledge, a major challenge is to capture value. Associated with this are a wide range of new organizational models. Distributed collaboration across organizations and time zones is increasingly common. The motivation for such collaboration is the desire to access sources of knowledge and skills not usually available in one place. The result of such changes has been a paradigm shift that has prompted the reengineering of organizations; the development of high-performance business teams, integrated organizations, and extended virtual enterprises; as well as the emergence of loosely structured organizations that have enhanced productivity. *See* DATA COMMUNICATIONS; INTERNET; LOCAL-AREA NETWORKS; WIDE-AREA NETWORKS.

Organizational challenges. There are a number of substantial challenges associated with use of information technology with respect to enhancing the productive efforts of an individual, a group, or an organization. It has been suggested that the command and control model of leadership is poorly suited to the management of organizations where the participants are not bound by traditional incentive and reward systems. A collaborative effort has to continue to make sense and to provide value to participants for it to be sustained. Otherwise, knowledge and skills are quite portable, and the loss of knowledge workers poses a major potential risk for organizations.

Decrease in costs. The information technology revolution is driven by technology and market considerations. Information technology costs declined in the 1990s due to the use of such technologies as broadband fiber optics, spectrum management, and data compression. The power of computers continues to increase; the cost of computing declined by a factor of 10,000 during from 1975–2000. Large central mainframe computers have been augmented, and in many cases replaced, by smaller, powerful, and more user-friendly personal computers. There has, in effect, been a merger of the computer and telecommunications industries into the information technology industry, and it now is possible to store, manipulate, process, and transmit voice, digitized data, and images at very little cost.

Benefits. As a consequence of the information technology revolution, information and knowledge have become powerful factors for socioeconomic development on a global scale. This knowledge has the potential to provide comprehensive support for enhanced production of goods and services, educational and employment opportunities for all peoples, institutions and infrastructures that enable better management in the private sector and governance in the public sector, natural resource conservation and environmental preservation, and global sustainable development. There are a number of complex issues that require attention, such as the impact of the information technology revolution on social, ethical, cultural, and family values. Most importantly, the benefits of the information technology revolution are overwhelmingly large and lead to new opportunities for world progress through the dissemination of sustainable development knowledge. Indeed, perhaps sustainable human development is not a realistic possibility at this time without major reliance on the support provided by information technology.

Andrew P. Sage

Bibliography. B. H. Boar, *Aligning Information Technology with Business Strategy*, John Wiley, New York, 1994; S. P. Bradley and R. L. Nolan, *Sense and Respond: Capturing Value in the Network Age*, Harvard Business School Press, Boston, 1998; F. Cairncross, *The Death of Distance: How the Communications Revolution Will Change Our Lives*, Harvard Business School Press, Boston, 1997; M. Dertouzos, *What Will Be: How the New World of Information Will Change Our Lives*, Harper Collins, San Francisco, 1997; P. F. Drucker, *Managing in a Time of Great Change*, Dutton, New York, 1995; R. K. Lester, *The Productive Edge: How U. S. Industries are Pointing the Way to a New Era of Economic Growth*, W. W. Norton, New York, 1998; D. Tapscott and A. Caston, *Paradigm Shift: The New Promise of Information Technology*, McGraw-Hill, New York, 1993.

Information theory

A branch of communication theory devoted to problems in coding. A unique feature of information theory is its use of a numerical measure of the amount of information gained when the contents of a message are learned. Information theory relies heavily on the mathematical science of probability. For this reason the term information theory is often applied loosely to other probabilistic studies in communication theory, such as signal detection, random noise, and prediction. *See* ELECTRICAL COMMUNICATIONS.

Information theory provides criteria for comparing different communication systems. The need for comparisons became evident during the 1940s. A large variety of systems had been invented to make better use of existing wires and radio spectrum allocations. In the 1920s the problem of comparing telegraph systems attracted H. Nyquist and R. V. L. Hartley, who provided some of the philosophy behind information theory. In 1948 C. E. Shannon published a precise general theory which formed the basis for subsequent work in information theory. *See* MODULATION; RADIO SPECTRUM ALLOCATION.

In information theory, communication systems are compared on the basis of signaling rate. Finding an appropriate definition of signaling rate was itself a problem, which Shannon solved by the use of his measure of information, to be explained later. Of special interest are optimal systems which, for a given set of communication facilities, attain a maximum signaling rate. Optimal systems provide communication-systems designers with useful absolute bounds on obtainable signaling rates. Although optimal systems often use complicated and expensive encoding equipment, they provide insight into the design of fast practical systems.

Communication systems. In designing a one-way communication system from the standpoint of information theory, three parts are considered beyond the control of the system designer: (1) the source, which generates messages at the transmitting end of the system, (2) the destination, which ultimately receives the messages, and (3) the channel, consisting of a transmission medium or device for conveying signals from the source to the destination. Constraints beyond the mere physical properties of the transmission medium influence the design. For example, in designing a radio system only a given portion of the radio-frequency spectrum may be available. The transmitter power may also be limited. If the system is just one link in a larger system which plans to use regenerative repeaters, the designer may be restricted to pulse-transmission schemes. All such conditions are considered part of the description of the channel. The source does not usually produce messages in a form acceptable as input by the channel. The transmitting end of the system contains another device, called an encoder, which prepares the source's messages for input to the channel. Similarly the receiving end of the system will contain a decoder to convert the output of the channel into a form recognizable by the destination. The encoder and decoder are the parts to be designed. In radio systems this design is essentially the choice of a modulator and a detector.

Discrete and continuous cases. A source is called discrete if its messages are sequences of elements (letters) taken from an enumerable set of possibilities (alphabet). Thus sources producing integer data or written English are discrete. Sources which are not discrete are called continuous, for example, speech and music sources. Likewise, channels are classified as discrete or continuous according to the kinds of signals they transmit. Most transmission media (such as transmission lines and radio paths) can provide continuous channels; however, constraints (such as a restriction to use pulse techniques) on the use of these media may convert them into discrete channels.

The treatment of continuous cases is sometimes simplified by noting that a signal of finite bandwidth can be encoded into a discrete sequence of numbers. If the power spectrum of a signal $s(t)$ is confined to the band zero to W hertz (cycles per second) then Eq. (1) applies. Equation (1) reconstructs

$$s(t) = \sum_{n=-\infty}^{\infty} s\left(\frac{n}{2W}\right) \frac{\sin 2\pi W\left(t - \dfrac{n}{2W}\right)}{2\pi W\left(t - \dfrac{n}{2W}\right)} \qquad (1)$$

$s(t)$ exactly from its sample values (Nyquist samples), at discrete instants $(2W)^{-1}$ s apart. Thus, a continuous channel which transmits such signals resembles a discrete channel which transmits Nyquist samples drawn from a large finite set of signal levels and at the rate of $2W$ samples per second.

Noiseless and noisy cases. The output of a channel need not agree with its input. For example, a channel might, for secrecy purposes, contain a cryptographic device to scramble the message. Still, if the output of the channel can be computed knowing just the input message, then the channel is called noiseless. If, however, random agents make the output unpredictable even when the input is known, then the channel is called noisy. *See* COMMUNICATIONS SCRAMBLING; CRYPTOGRAPHY.

Encoding and decoding. Many encoders first break the message into a sequence of elementary blocks; next they substitute for each block a representative code, or signal, suitable for input to the channel. Such encoders are called block encoders. For example, telegraph and teletype systems both use block encoders in which the blocks are individual letters. Entire words form the blocks of some commercial cablegram systems. The operation of a block encoder may be described completely by a function or table showing, for each possible block, the code that represents it.

It is generally impossible for a decoder to reconstruct with certainty a message received via a noisy channel. Suitable encoding, however, may make the noise tolerable, as may be illustrated by a channel that transmits pulses of two kinds. It is customary to let binary digits 0 and 1 denote the two kinds of pulse. If the source has only the four letters A, B, C, D, it is possible to simply encode each single-letter block into a pair of binary digits (code I; see **table**). In that case the decoder would make a mistake every time that noise produced an error. If each single-letter block is encoded into three digits (code II), the decoder can at least recognize that a received triple of digits must contain errors if it is one of the triples (001, 010, 100, or 111) not listed in the code. Because an error in any one of the three pulses of code II always produces a triple that is not listed, code II provides single-error detection. Similarly, a five-binary-digit code (code III) can provide double-error detection, because errors in a single pulse or pair of pulses always produce a quintuple that is not listed.

Three possible binary codes for four-letter alphabet

Letter	Code I	Code II	Code III
A	0 0	0 0 0	0 0 0 0 0
B	0 1	0 1 1	0 0 1 1 1
C	1 0	1 0 1	1 1 0 0 1
D	1 1	1 1 0	1 1 1 1 0

As an alternative, code III may provide single-error correction. In this usage, the decoder picks a letter for which code III agrees with the received quintuple in as many places as possible. If only a single digit is in error, this rule chooses the correct letter.

Even when the channel is noiseless, a variety of encoding schemes exists and there is a problem of picking a good one. Of all encodings of English letters into dots and dashes, the Continental Morse encoding is nearly the fastest possible one. It achieves its speed by associating short codes with the most common letters. A noiseless binary channel (capable of transmitting two kinds of pulse 0, 1, of the same duration) provides the following example. In order to encode English text for this channel, a simple encoding might just use 27 different five-digit codes to represent word space (denoted by #), A, B, . . . , Z; say # 00000, A 00001, B 00010, C 00011, . . . , Z 11011. The word #CAB would then be encoded into 00000000110000100010. A similar encoding is used in teletype transmission; however, it places a third kind of pulse at the beginning of each code to help the decoder stay in synchronism with the encoder. The five-digit encoding can be improved by assigning four-digit codes 0000, 0001, 0010, 0011, 0100 to the five most common letters #, E, T, A, O. There are 22 quintuples of binary digits which do not begin with any of the five four-digit codes; these may be assigned as codes to the 22 remaining letters. About half the letters of English text are #, E, T, A, or O; therefore the new encoding uses an average of only 4.5 digits per letter of message. *See* TELETYPEWRITER.

More generally, if an alphabet is encoded in single-letter blocks, using $L(i)$ digits for the ith letter, the average number of digits used per letter is shown in Eq. (2), where $p(i)$ is the probability of the ith letter.

$$L = p(1)L(1) + p(2)L(2) + p(3)L(3) + \cdots \qquad (2)$$

An optimal encoding scheme will minimize L. However, the encoded messages must be decipherable, and this condition puts constraints on the $L(i)$. The code lengths of decipherable encodings must satisfy the relationships shown in inequality (3). The real

$$2^{-L(1)} + 2^{-L(2)} + 2^{-L(3)} + \cdots \leq 1 \qquad (3)$$

numbers $L(1)$, $L(2), \ldots$, which minimize L subject to inequality (3) are $L(i) = -\log_2 p(i)$ and the corresponding minimum L is shown in Eq. (4), which

$$H = -\sum_i p(i) \log_2 p(i) \qquad (4)$$

provides a value of H equal to a number of digits per letter.

The $L(i)$ must be integers and $-\log_2 p(i)$ generally are not integers; for this reason there may be no encoding which provides $L = H$. However, Shannon showed that it is always possible to assign codes to letters in such a way that $L \leq H + 1$. A procedure for constructing an encoding which actually minimizes L has been given by D. A. Huffman. For (27-letter) English text $H = 4.08$ digits per letter, as compared with the actual minimum 4.12 digits per letter obtained by Huffman's procedure.

By encoding in blocks of more than one letter, the average number of digits used per letter may be reduced further. If messages are constructed by picking letters independently with the probabilities $p(1)$, $p(2), \ldots$, then H is found to be the minimum of the average numbers of digits per letter used to encode these messages using longer blocks. *See* DATA COMPRESSION.

Information content of message. The information contained in a message unit is defined in terms of the average number of digits required to encode it. Accordingly the information associated with a single letter produced by a discrete source is defined to be the number H. Some other properties of H help to justify using it to measure information. If one of the $p(i)$ equals unity, only one letter appears in the messages. Then nothing new is learned by seeing a letter and, indeed, $H = 0$. Second, of all possible ways of assigning probabilities $p(i)$ to an N-letter alphabet, the one which maximizes H is $p(1) = p(2) = \cdots = 1/N$. This situation is the one in which the unknown letter seems most uncertain; therefore it does seem correct that learning such a letter provides the most information. The corresponding maximum value of H is $\log_2 N$. This result seems reasonable by the following argument. When two independent letters are learned, the information obtained should be $2H = 2 \log_2 N$. However, such pairs of letters may be considered to be the letters of a larger alphabet of N^2 equally likely pairs. The information associated with one of these new letters is $\log_2 N^2 = 2 \log_2 N$. Although H given by Eq. (4) is dimensionless, it is given units called bits (a contraction of binary digits). Occasionally the information is expressed in digits of other kinds (such as ternary or decimal). Then bases other than 2 are used for the logarithm in Eq. (4).

The majority of message sources do not merely pick successive letters independently. For example in English, H is the most likely letter to follow T but is otherwise not common. The source is imagined to be a random process in which the letter probabilities change, depending on what the past of the message has been. Statistical correlations between different parts of the message may be exploited by encoding longer blocks. The average number of digits per letter may thereby be reduced below the single-letter information H given by Eq. (4). For example, by encoding English words instead of single letters, 2.1 digits/letter suffice. Encoding longer and longer blocks, the number of digits needed per letter approaches a limiting minimum value. This limit is called the entropy of the source and is interpreted as the rate, in bits per letter, at which the source generates information. If the source produces letters at some fixed average rate, n letters/s, the entropy may also be converted into a rate in bits per second by multiplying by n. The entropy may be computed from tables giving the probabilities of blocks of N letters (N-grams). If in Eq. (4) the summation index i is extended over all N-grams, then the number H represents the information in N consecutive letters. As $N \to \infty$, H/N approaches the entropy of the source. The entropy of English has been estimated by Shannon to be about 1 bit/letter. However, an encoder might have to encode 100-grams to achieve a reduction to near 1 digit/letter. Comparing English with a source that produces 27 equally likely letters independently (hence has entropy $\log_2 27 = 4.8$ bits/letter), this result is often restated: English is 80% redundant. *See* ENTROPY.

In computer applications, coding can reduce the redundancy of data for storage, saving space in memory. Large white areas make black-and-white diagrams very redundant. To store diagrams or send them by facsimile, coding white runs instead of individual black-and-white samples will remove much redundancy. *See* FACSIMILE.

Universal codes. Even when a probabilistic description of the source is unavailable, coding can remove redundancy. In a simple illustration, the encoder begins by parsing a binary message into blocks, now of varying lengths. Each new block B is made as long as possible, but B must be a block B', seen earlier, followed by one extra digit x. A typical parsed message is 0,01,00,1,000, . . . , with commas separating the blocks. When transmitting the kth block B, the transmitter first sends the last digit x of B. It next sends a string of digits, of length about $\log_2 k$, to identify the earlier block B'. Redundancy makes block sizes tend to grow faster than $\log_2 k$; then the encoding reduces message length. A. Lempel and J. Ziv have shown that codes like this one compress data about as well as codes that use given source statistics. In effect, the parsing gathers statistics to help encode future blocks.

Capacity. The notion of entropy is more widely applicable than might appear from the discussion of the binary channel. Any discrete noiseless channel may be given a number C, which is called the capacity. C is defined as the maximum rate (bits per second) of all sources that may be connected directly to the channel. Shannon proved that any given source (which perhaps cannot be connected directly to the channel) of entropy H bits/letter can be encoded for the channel and run at rates arbitrarily close to C/H letters/s.

By using repetition, error-correcting codes, or similar techniques, the reliability of transmission over a noisy channel can be increased at the expense of slowing down the source. It might be expected that the source rate must be slowed to 0 bits/s as the transmission is required to be increasingly error-free. On the contrary, Shannon proved that even a noisy channel has a capacity C. Suppose that errors in at most a fraction ϵ of the letters of the message can be tolerated ($\epsilon > 0$). Suppose also that a given source, of entropy H bits/letter, must be operated at the rate of at least $(C/H) - \delta$ letters/s ($\delta > 0$). No matter how small ϵ and δ are chosen, an encoder can be found which satisfies these requirements.

For example, the symmetric binary channel has binary input and output letters; noise changes a fraction p of the 0's to 1 and a fraction p of the 1's to 0 and treats successive digits independently. The capacity of this channel is shown by Eq. (5), where m is

$$C = m[1 + p \log_2 p + (1 - p) \log_2(1 - p)] \qquad (5)$$

the number of digits per second which the channel transmits.

Shannon's formula, shown by Eq. (6), gives the ca-

$$C = W \log_2 \left(1 + \frac{S}{N}\right) \qquad (6)$$

pacity C of a band-limited continuous channel. The channel consists of a frequency band W Hz wide, which contains a gaussian noise of power N. The noise has a flat spectrum over the band and is added to the signal by the channel. The channel also contains a restriction that the average signal power may not exceed S.

Equation (6) illustrates an exchange relationship between bandwidth W and signal-to-noise ratio S/N. By suitable encoding a signaling system can use a smaller bandwidth, provided that the signal power is also raised enough to keep C fixed. *See* BANDWIDTH REQUIREMENTS (COMMUNICATIONS).

Typical capacity values are 20,000 bits/s for a telephone speech circuit and 50,000,000 bits/s for a broadcast television circuit. Speech and television are very redundant and would use channels of much lower capacity if the necessary encodings were inexpensive. For example, the vocoder can send speech, only slightly distorted, over a 2000-bits/s channel. Successive lines or frames in television tend to look alike. This resemblance suggests a high redundancy; however, to exploit it the encoder may have to encode in very long blocks.

Not all of the waste in channel capacity can be attributed to source redundancies. Even with an irredundant source, such as a source producing random digits, some channel capacity will be wasted. The simplest encoding schemes provide reliable transmission only at a rate equal to the capacity of a channel with roughly 8 dB smaller signal power (the 8-dB figure is merely typical and really depends on the reliability requirements). Again, more efficient encoding to combat noise generally requires larger-sized blocks. This is to be expected. The signal is separated from the noise on the basis of differences between the signal's statistical properties and those of noise. The block size must be large enough to supply the decoder with enough data to draw statistically significant conclusions. *See* ELECTRICAL NOISE.

Algebraic codes. Practical codes must use simple encoding and decoding equipment. Error-correcting codes for binary channels have been designed to use small digital logic circuits. These are called algebraic codes, linear codes, or group codes because they are constructed by algebraic techniques involving linear vector spaces or groups.

For example, each of the binary codes I, II, and III discussed above (see table) contains four code words which may be regarded as vectors $C = (c_1, c_2, \ldots, c_n)$ of binary digits c_i. The sum $C + C'$ of two vectors may be defined to be the vector $(c_1 + c'_n, \ldots, c_n + c'_n)$ in which coordinates of C and C' are added modulo 2. Codes I, II, and III each have the property that the vector sum of any two code words is also a code word. Because of that, these codes are linear vector spaces and groups under vector addition. Their code words also belong to the n-dimensional space consisting of all 2^n vectors of n binary coordinates. Codes II and III, with $n = 3$ and 5, do not contain all 2^n vectors; they are only two-dimensional linear subspaces of the larger space. Consequently, in Codes II and III, the coordinates c_i must satisfy certain linear homogeneous equations. Code II satisfies $c_1 + c_2 + c_3 = 0$. Code III satisfies $c_3 + c_4 = 0$, $c_2 + c_3 + c_5 = 0$, $c_1 + c_2 = 0$, and other equations linearly dependent on these three. The sums in such equations are performed modulo 2; for this reason the equations are called parity check equations. In general, any r linearly independent parity check equations in $c_1, \ldots, c_n$ determine a linear subspace of dimension $k = n - r$. The 2^k vectors in this subspace are the code words of a linear code. *See* GROUP THEORY; LINEAR ALGEBRA; NUMBER THEORY.

The r parity checks may be transformed into a form which simplifies the encoding. This transformation consists of solving the original parity check equations for some r of the coordinates c_i as expressions in which only the remaining $n - r$ coordinates appear as independent variables. For example, the three parity check equations given for Code III are already in solved form with c_1, c_4, c_5 expressed in terms of c_2 and c_3. The $k = n - r$ independent variables are called message digits because the 2^k values of these coordinates may be used to represent the letters of the message alphabet. The r dependent coordinates, called check digits, are then easily computed by circuits which perform additions modulo 2. *See* LINEAR SYSTEMS OF EQUATIONS.

At the receiver the decoder can also do additions modulo 2 to test if the received digits still satisfy the parity check equations. The set of parity check equations that fail is called the syndrome because it contains the data that the decoder needs to diagnose the errors. The syndrome depends only on the error locations, not on which code word was sent. In general, a code can be used to correct e errors if each pair of distinct code words differ in at least $2e + 1$ of the n coordinates. For a linear code, that is equivalent to requiring the smallest number d of "ones" among the coordinates of any code word [excepting the zero word $(0, 0, \ldots, 0)$] to be $2e + 1$ or more. Under these conditions each pattern of $0, 1, \ldots, e - 1$, or e errors produces a distinct syndrome; then the decoder can compute the error locations from the syndrome. This computation may offer some difficulty. But at least it involves only r binary variables, representing the syndrome, instead of all n coordinates.

Hamming codes. The r parity check equations may be written concisely as binary matrix equation (7). Here C_T is a column vector, the transpose of

$$HC^T = 0 \tag{7}$$

$(c_1, \ldots, c_n)$. H is the so-called parity check matrix, having n columns and r rows. A Hamming single-error correcting code is obtained when the columns of H are all $n = 2^r - 1$ distinct columns of r binary digits, excluding the column of all zeros. If a single error occurs, say in coordinate c_i, then the decoder uses the syndrome to identify c_i as the unique coordinate that appears in just those parity check equations that fail. *See* MATRIX THEORY.

Shift register codes. A linear shift register sequence is a periodic infinite binary sequence $\ldots, c_0, c_1, c_2, \ldots$ satisfying a recurrence equation expressing c_j as a modulo 2 sum of some of the b earlier digits $c_{j-b}, \ldots, c_{j-1}$. A recurrence with two terms would be an equation $c_j = c_{j-a} + c_{j-b}$, with a equal to some integer $1, 2, \ldots$, or $b - 1$. The digits of a shift register sequence can be computed, one at a time, by very simple equipment. It consists of a feedback loop, containing a shift register to store $c_{j-b}, \ldots, c_{j-1}$ and a logic circuit performing modulo 2 additions. This equipment may be used to implement a linear code. First, message digits $c_1, \ldots, c_b$ are stored in the register and transmitted. Thereafter the equipment computes and transmits successively the $n - b$ check digits c_j obtained from the recurrence equation with $j = b + 1, \ldots, n$. By choosing a suitable recurrence equation, the period of the shift register sequence can be made as large as $2^b - 1$. Then, with n equal to the period $2^b - 1$, the code consists of the zero code word $(0, 0, \ldots, 0)$ and $2^b - 1$ other code words which differ from each other only by cyclic permutations of their coordinates. These latter words all contain $d = 2^b - 1$ "ones" and so the code can correct $e = 2^{b-2} - 1$ errors. *See* SWITCHING CIRCUIT.

Intermediate codes. The Hamming codes and maximal period shift register codes are opposite extremes, correcting either one or many errors and having code words consisting either mostly of message digits or mostly of check digits. Many intermediate codes have been invented. One of them requires $n + 1$ to be a power of 2; say $n + 1 = 2^q$. It then uses at most qe check digits to correct e errors.

Perfect codes. Although each pattern of $0, 1, \ldots, e$ errors produces a distinct syndrome, there may be extra syndromes which occur only after more than e errors. In order to keep the number of check digits small, extra syndromes must be avoided. A code is called perfect if all 2^r syndromes can result from patterns of $0, 1, \ldots, e - 1$, or e errors. Hamming codes are all perfect. M. J. E. Golay found another perfect binary code having $n = 23$, $r = 11$ check digits, and correcting $e = 3$ errors.

Orthogonal parity codes. Orthogonal parity codes are codes with especially simple decoding circuits which take a kind of majority vote. Suppose the parity check equations can be used to derive $2e + 1$ linear equations in which one digit, say c_1, is expressed with each of the remaining digits $c_2, \ldots, c_n$ appearing in at most one equation. If at most e errors occur, then the received digits satisfy a majority of the $2e + 1$ equations if and only if c_1 was received correctly. For example, the recurrence $c_j = c_{j-2} + c_{j-3}$ generates a maximal period shift register code with $n = 7$, $r = 4$, $e = 1$. With $j = 1$, 3, and 4 in the recurrence equation, three of the parity check equations, $c_1 = c_5 + c_6$, $c_1 = c_3 + c_7$, and $c_1 = c_2 + c_4$, are obtained, after using the fact that the shift register sequence has period 7. These three equations are already in the form required for decoding c_1 by majority vote. Similar equations, obtained by permuting $c_1, \ldots, c_7$ cyclically, apply for $c_2, \ldots, c_7$. Then the decoder can be organized so that most of the equipment used to decode c_1 can be used again in decoding $c_2, \ldots, c_7$. *See* COMBINATORIAL THEORY.

Wide-band signaling. A band-limited continuous channel has capacity C, given by Eq. (6), very near $1.44\ WS/N$ if the signal-to-noise ratio S/N is small. Then high capacity is obtained even with low signal power S if the bandwidth W is wide enough. Wide bands are used in satellite systems and in military applications where a weak signal must hide in background noise or penetrate jamming. Or, many users can share the same wide radio channel effectively even though their signals spread over the entire band and interfere with one another.

One method of wide-band signaling modulates a kind of noise carrier, a stream of short pulses of random sign, by changing signs in accordance with the digits of a binary message. The receiver has a synchronous detector that must regenerate the carrier locally in order to demodulate the signal. A carrier kept secret from other listeners can serve as the running cryptographic key of a stream cipher. Noiselike carriers may take the form of long periodic sequences (for example, Gold codes) generated by linear recurrences in the manner of shift-register codes. *See* AMPLITUDE-MODULATION DETECTOR; SPREAD SPECTRUM COMMUNICATION.

Nonblock codes. Many useful codes do not send messages block by block. Simple convolutional codes intersperse check digits among message digits in a regular pattern. An example, designed to correct errors that cluster together in isolated bursts, transmits the binary message $\ldots, x_1, x_2, x_3, \ldots$ as $\ldots, c_1, x_1, c_2, x_2, c_3, x_3, \ldots$, with $c_i = x_{i-3} + x_{i-6}$ (modulo 2).

It can correct any burst of length 6 or less that lies more than 19 digits away from other errors.

Convolutional codes are commonly decoded by sequential or dynamic programming methods such as the Viterbi algorithm. Using the given probabilistic description of the noise process, the decoder calculates the posterior probabilities of messages that might possibly have produced the data received up to the present time. The decoder then chooses the most likely message as its decoded message. The calculation is done digit by digit, using the probabilities computed for the first k received digits to simplify the calculation for $k + 1$ digits. Although the number of possible messages grows very quickly with k, one can eliminate highly unlikely messages to keep the calculation reasonably short at each step. *See* ESTIMATION THEORY.

For signaling over radio channels with additive gaussian noise, trellis codes, which are particular kinds of nonblock codes, often have an advantage over block codes of equal complexity. Expressed in terms of signal-to-noise ratio, the advantage can be 3 dB or more even for simple codes. Edgar N. Gilbert

Bibliography. E. R. Berlekamp, *Algebraic Coding Theory*, rev. ed., 1984; R. W. Hamming, *Coding and Information Theory*, 2d ed., 1986; R. W. Lucky, *Silicon Dreams: Information, Man, and Machine*, 1991; M. Schwartz, *Information Transmission, Modulation, and Noise*, 4th ed., 1990; N. J. A. Sloane and F. J. MacWilliams, *Theory of Error-Correcting Codes*, 9th reprint, 1998; C. E. Shannon, *Collected Papers*, 1993.

Information theory (biology)

The application to biological systems of the theory of encoding effects on signal transmission and communications efficiency. In everyday language, information is associated with knowledge, meaning (semantics), and the influence of information on behavior (pragmatics). Information theory considers information to be quantitative and capable of being expressed in binary digits, or bits. The measure of information is based on its structure or its representation, taking into account the statistical properties of its structure. Mathematical information theory was created to describe the transmission of information in telecommunications, but it has been widely applied also in mathematics, physics, and biology.

Representation of information. A physical quantity that carries information is called a signal. A signal may be represented in space and time. Examples include speech and radio waves, time-dependent voltage, a still or moving picture, and the pulse sequence within a nerve fiber. In order to calculate the amount of information it contains, a signal must be quantized in a two-step process for symbolic presentation. First, the time or space coordinate has to be discrete, appearing only at precise, usually equidistant points on the coordinates. For signals with limited variations (limited bandwidth), this is possible without a loss of generality because of the sampling theorem. Second, the amplitude of the signal must be quantized, or expressed in (usually equal) steps. The size of the steps, which is arbitrary, determines the accuracy of the representation, but it is usually chosen at the limit of resolution to detect a certain variation of the amplitude. The signal can then be represented by a finite number of discrete points in time or space and by amplitude levels at each point.

An information source is a device that produces signals that carry information. In the symbolic representation of information theory, these signals are discrete in time or space and quantized in amplitude, and the information source is usually described by a stochastic process. Written language is a typical example of the signal sequence produced by a stationary information source.

Coding theorem of information. The coding theorem of information shows that it is possible to code the output of an information source with a number of bits equal to its entropy, or uncertainty with respect to the signals of the source. Symbols or symbol sequences with a high probability are represented by a short code word, and those with low probability are represented by a longer code word, much like the Morse alphabet.

The coding theorem of information restricts the complexity needed to represent the information produced by an information source, and that restriction is important in telecommunications. Source coding makes use of the redundancy of the information source to reduce the amount of data to be transmitted. In the case of speech transmission with telephone quality, a reduction factor of about 1:10 has been achieved. For a television or picture phone signal, reduction factors of 1:100 or even 1:1000 are hypothesized. *See* ENTROPY.

Information channel. The information channel is the connection of an information source with a receiver. The information source is connected by way of a coder to the channel. The coder transforms the signals of the source into transmissible signals, eliminates redundancy, and reduces the effect of noise introduced by the channel (channel coding). Suitable modulation may be included, and when performed simultaneously with channel coding, is called codulation. The transmission channel may have frequency dispersion (channel with memory). The channel typically introduces noise, which in the simplest case is additive to the signal and has a gaussian amplitude distribution, but the scheme also holds for more general cases, such as multiplicative noise and Poisson distributions. The decoder delivers the decoded signal to the receiver or destination. Entropy at the input and entropy at the output of the channel can then be calculated: if only noise without dispersion is present, perfect transmission without errors or loss of information is possible.

Coding theorem of transinformation. The coding theorem of transinformation states that in the limiting case of very long sequences a transmission with vanishing error probability is possible. The ideal case may be approximated by coding very long signal sequences, which also implies a very long delay between transmitter and receiver. However, error-free transmission is possible even in noisy surroundings. Error-correcting codes can approach the ideal error-free case, and many optimized binary coding schemes exist.

The decoder, which can be considered equivalent to a pattern recognition device, may have a very high complexity, but it can be reduced by algorithms. Its task is to classify or detect the desired signal in a set of possible received signals. The human receiver, or decoder, has the same task of pattern recognition, which is performed by the nervous system. *See* CYBERNETICS.

Human information channel. Information theory can be applied to human information processing (**Fig. 1**). The number of receptors for each of various sensory organs has been determined, and the numerical values of maximum information flow for some of the sensory organs has been estimated. The data given in Fig. 1 show that television and telephone channels are well adapted to human receptors. *See* INFORMATION PROCESSING (PSYCHOLOGY).

Information flow for various conscious human activities can also be estimated. Counting requires 3 bits/s; reading, 40 bits/s; hearing, 40 bits/s; and piano playing, 20 bits/s. The maximum human channel capacity for conscious output is 50 bits/s, which indicates an apparent reduction of information in the central nervous system. Such reduction is necessary because not all details of the external world need to penetrate the consciousness. A good deal of this information is superfluous or redundant and therefore can be eliminated; otherwise memory would be unable to store the entire flow of sensory information. If 50 bits/s is multiplied by a life expectancy of 70 years, the information received totals about 10^{10} bits. The storage capacity of human memory is estimated to be of that order of magnitude. *See* MEMORY.

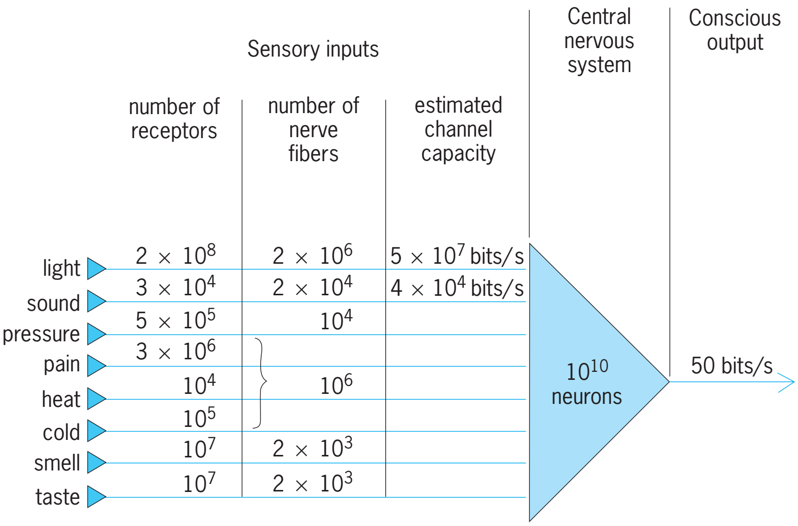


Fig. 1. Human information processing. The estimates of channel capacity are based on the number of rods in the retina of the eye and the number of hair cells in the inner ear's organ of Corti.

Controlled source and bidirectional communication. In human communication, a unidirected information channel is a monologue. The usual situation, however, is a dialogue, which involves bidirectional communication. In bidirectional communication theory, the independent information source is replaced by a controlled or dependent information source. In a dialogue, the transmitter and the receiver are both described by controlled information sources. The behavior of the living being may be described statistically by a dependent information source, in which controlling input represents the sensory signals and the output is the behavior. The mathematical description of a controlled information source is given by the conditional probability $p(x|x_n y_n)$ where x is the produced symbol, x_n sequence of previously produced symbols, and y_n the sequence of previously received symbols. The expression $p(x|x_n y_n)$ also represents generator probability independent of the combination of the transmitted and received sequences $x_n y_n$. If behavior changes because of a learning or adaptive process, the probability changes.

For the dependent information source, two entropy quantities may be defined. The first is the usual entropy, given by Eq. (1). This takes into account

$$H(x) = \lim_{n\to\infty} E[-\log p(x|x_n)] \qquad (1)$$

only the produced signal sequence x_n. If both the produced and the received signal sequences x_n and y_n are considered, the free entropy is determined by Eq. (2).

$$F(x) = \lim_{n\to\infty} E[-\log p(x|x_n y_n)] \qquad (2)$$

Clearly $F(x) \le H(x)$ because it contains the knowledge of the received signals. The difference between the entropies, which is known as directed transinformation $y \to x$, is given by Eq. (3), where $T(x|y)$ represents the influence of the controlling (or received) signals on the information production of the source,

$$T(x|y) = H(x) - F(x) \qquad (3)$$

and is thus the transinformation from the y sequence to the x sequence. The whole entropy $H(x)$ of the source thus consists of a free part $F(x)$ and a dependent part $T(x|y)$ transmitted by the y sequence.

A bidirectional communication takes place when two dependent information sources are coupled. The entropy quantities for that case are shown in the **table**. The sum of the two directed transinformation quantities equals the transinformation calculated by the original information theory. If one of the two transinformations, $T(x|y)$ or $T(y|x)$, is zero, the problem is reduced to the special case of the unidirectional information channel. The two equations for the directed transinformation lead to the flow diagram of the communication shown in **Fig. 2**. This diagram also shows that only part of the produced entropy $H(x)$, namely $T(y|x)$, reaches the counterpart. Thus the coupling may be small because of noise or other limitations of the communication channels.

Quantities associated with entropy of bidirectional communication

Quantity	Equation
Entropy of *x*	$H(x) = \lim_{n\to\infty} E[-\log p(x/x_n)]$
Entropy of *y*	$H(y) = \lim_{n\to\infty} E[-\log p(y/y_n)]$
Free entropy of *x*	$F(y) = \lim E[-\log p(y/x_n y_n)]$
Free entropy of *y*	$F(y) = \lim_{n\to\infty} E[-\log p(y/x_n y_n)]$
Directed transinformation $y \longrightarrow x$	$T(x/y) = H(x) - F(x)$
Directed transinformation $x \longrightarrow y$	$T(y/x) = H(y) - F(y)$
Total transinformation (coincidence, synentropy) $x \longleftrightarrow y$	$T = T(x/y) + T(y/x)$

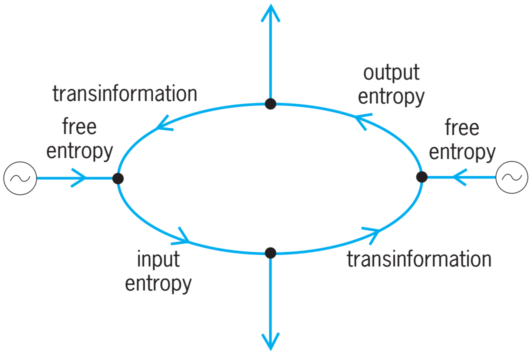


Fig. 2. Flow diagram of communication.

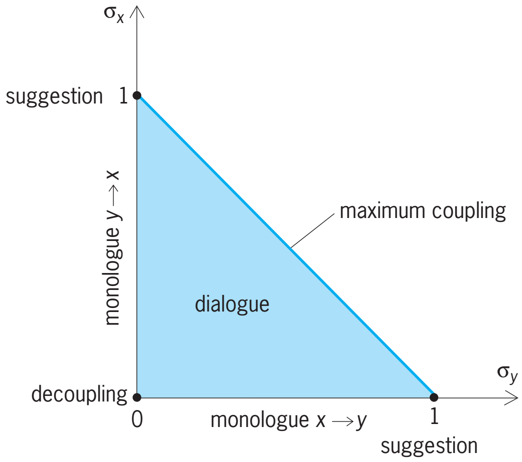


Fig. 3. Coupling diagram of communication.

Maximal coupling, however, is limited. If the coupling coefficients G_x and G_y are defined by Eqs. (4) and (5), then the limiting condition $\sigma_x + \sigma_n \leq 1$ holds.

$$G_x = \frac{T(x|y)}{H(x)} \qquad (4)$$

$$G_y = \frac{T(y|x)}{H(y)} \qquad (5)$$

According to that inequality, all possible coupling states of a communication are shown in the coupling diagram of **Fig. 3**. The maximum coupling occurs when $\sigma_x + \sigma_y = 1$. Complete (quasi-deterministic) coupling can obviously occur in only one direction. With respect to a human communication, it is referred to as suggestion. Bidirectional communication theory has been applied to the group behavior of monkeys, but it may also be used to describe the theoretical aspects of user-machine interaction. *See* HUMAN-MACHINE SYSTEMS; INFORMATION THEORY.

Hans Marko

Bibliography. R. E. Blahut, *Communications, Probability, Information Theory, Coding*, 1987; C. Cherry, *On Human Communication: A Review, a Survey and a Criticism*, 3d ed., 1978; R. G. Gallager, *Information Theory and Reliable Communication*, 1968; D. L. Kincaid (ed.), *Communication Theory: Eastern and Western Perspectives*, 1987; S. Oyama, *Ontogeny of Information: Developmental Systems and Evolution*, 2d ed., 2000; J. R. Sampson (ed.), *Biological Information Processing: Current Theory and Computer Simulation*, 1989; C. Sybsema, *Biophysics: An Introduction*, 1989.

Infrared astronomy

The field of astronomical observations specializing in detecting photons from the infrared portion of the electromagnetic spectrum. Detection technology has traditionally defined the domains of the electromagnetic spectrum. The infrared portion of the spectrum spans the range from the red limit of human vision (approximately 0.7 micrometer) to the shortest wavelengths accessible to heterodyne radio receivers (several hundred micrometers). *See* ELECTROMAGNETIC RADIATION; INFRARED RADIATION.

Differences from optical astronomy. Astronomers observing the universe with infrared light encounter a number of fundamental differences relative to those observing with visible light:

1. Infrared observations are more sensitive to cooler objects than visible-wavelength observations. Blackbodies at a temperature cooler than 2000 kelvins (3100°F) radiate virtually all of their light in the infrared part of the spectrum (**Fig. 1**). Infrared observations are particularly well suited to detect both forming stars and evolved stars (that is, stars in the final stages of their lives) since both classes of objects are cool. Since starlight often heats nearby dust grains to temperatures of tens or hundreds of degrees, reprocessing the visible starlight into exclusively infrared radiation, warm dust is also a common target for infrared observations. *See* HEAT RADIATION.

2. Interstellar dust is substantially more transparent at infrared wavelengths than at visible wavelengths. The dust grains constituting the interstellar medium tend to be smaller than 1 μm in diameter and are mainly composed of silicate and carbon grains. The theory of the scattering of electromagnetic radiation by dielectric particles dictates that such particles efficiently scatter wavelengths of light which

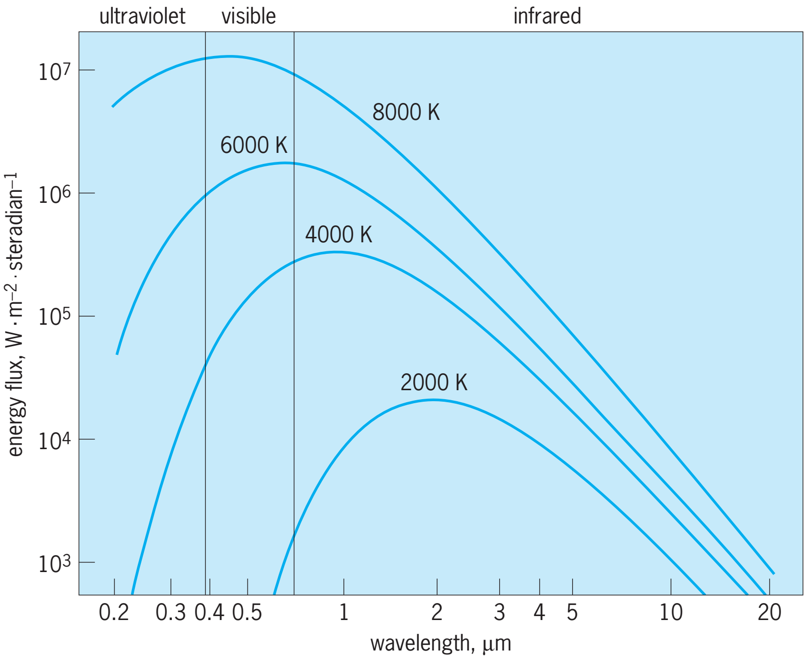


Fig. 1. Plots of the energy flux emitted by a blackbody as a function of wavelength for several temperatures. Hotter objects are substantially brighter and emit most of their energy at shorter wavelengths. Objects cooler than a few thousand kelvins emit predominantly infrared radiation. (The vertical axis is the product of the energy flux per unit wavelength and the wavelength itself.)

are equal to or smaller than the particle's size. Since infrared light has a wavelength substantially larger than the size of interstellar grains, the grains are ineffective at extinguishing the infrared light of stars and galaxies which lie behind the dust. Infrared observations thus enable astronomers to view distant objects through the obscuring dust that permeates the Milky Way Galaxy. Forming and evolved stars often reside in dense clouds of interstellar dust grains and can be observed only at infrared and even longer radio wavelengths. *See* INTERSTELLAR EXTINCTION; INTERSTELLAR MATTER; SCATTERING OF ELECTROMAGNETIC RADIATION.

3. The quantized energies of molecular rotational and vibrational transitions, which give rise to molecular spectral lines, fall largely in the infrared part of the spectrum, as do many hyperfine lines of individual atoms. Infrared spectral lines of water, molecular hydrogen, and carbon monoxide probe dense interstellar environments and cool stellar atmospheres. Very small interstellar grains can be considered as immense molecules. The bending and stretching modes of (CH) and (SiO) on the surface of these grains produce broad infrared spectral features diagnostic of the grains and, in cold environments such as dark interstellar clouds, of their icy mantles. *See* INFRARED SPECTROSCOPY.

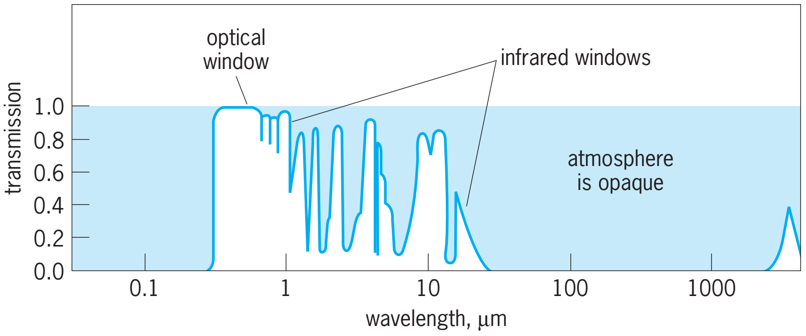


Fig. 2. Atmospheric transmission as a function of wavelength throughout the visible and infrared parts of the spectrum, showing atmospheric windows.

In addition to these astrophysical differences, the technology and practice of infrared astronomy differs from visible-wavelength astronomy in fundamental ways:

1. The Earth's atmosphere is opaque to infrared radiation through a substantial fraction of the spectrum (**Fig. 2**). This opacity arises largely from water in the Earth's lower atmosphere. Water's infrared spectral absorption lines blanket the infrared spectrum, particularly when blurred together by the pressure broadening which occurs in the dense terrestrial atmosphere. At the shorter infrared wavelengths, the infrared spectrum has several windows of transmission which permit ground-based telescopes to observe celestial sources. At the longer infrared wavelengths, the atmosphere is nearly opaque, and astronomers must resort to high-flying aircraft or orbiting satellites in order to evade the atmospheric opacity. *See* SATELLITE (ASTRONOMY).

2. The energy of chemical bonds and the work function for the liberation of electrons from a metal via the photoelectric effect is of order a few electronvolts. Planck's law, $E = h\nu = hc/\lambda$, which relates the energy E of a photon to its frequency ν or its wavelength λ via Planck's constant ($h = 6.63 \times 10^{-34}$ joule-second) and the speed of light ($c = 3 \times 10^8$ m/s), dictates that only photons with wavelength shorter than about 1 μm can induce chemical reactions or liberate free electrons. Thus physics limits photography and photon-counting photomultiplier tubes to operation mainly in the visible-wavelength domain. Infrared detection technology relies largely on either mimicking the photoelectric effect inside a crystalline semiconductor material or monitoring the temperature change of a semiconductor under the influence of infrared radiation. *See* PHOTOEMISSION.

3. Objects of room temperature (300 K or 80°F) emit radiation throughout most of the infrared spectrum. This glow interferes with the detection of faint astronomical sources, limiting the sensitivity of observations. Rapid comparison and differencing of the signal from the source direction versus the adjacent sky can mitigate some of these effects. Cooling of the entire telescope—impractical on the ground but possible in the vacuum of space—can substantially reduce the thermal glow of the telescope optics and result in unprecedented sensitivity to faint astronomical sources.

Infrared technology. Detectors of infrared radiation divide into two classes: bolometers and photovoltaic or photoconductive devices. Bolometer detectors have temperature-sensitive electrical conductivity. Incident radiation warms the detector, and the resulting subtle change in the electrical resistance of the detector is measured. Infrared photodetectors are crystalline semiconductors in which some electrons in the crystal lattice lie only a short distance in energy away from becoming unbound and behaving like metallic conducting electrons. Infrared light with energy in excess of the binding energy creates free charge carriers, either changing the bulk conductivity of the device (photoconductors) or charging or discharging a semiconductor junction (photovoltaics). The detector's response to infrared radiation depends on the chemistry-dependent solid-state physics of the detector material. In a mixture of indium and antimony (InSb), for example, the binding energy of the electrons corresponds to an infrared wavelength of 5.5 μm. This material is responsive to photons of this wavelength or shorter. Different semiconductor materials have different long-wavelength cutoffs. By doping the semiconductor material with impurities, it is possible to produce even more weakly bound electrons, yielding detectors with sensitivity to wavelengths as long as 100 μm or more. *See* BOLOMETER; PHOTOCONDUCTIVITY; PHOTOVOLTAIC EFFECT.

Since the mid-1980s, large-scale integration of semiconductor components has permitted the production of arrays of infrared detectors. These arrays now exist in formats as large as 2048 × 2048 elements (4 million detectors on a single device). Each detector on such an array is also substantially more sensitive than its 1980s counterpart. Infrared astronomy has blossomed thanks to the multiple advantages of observing with so many detectors simultaneously.

Systems of mirrors, gratings, prisms, and lenses direct light to infrared detectors in applications such as photometers, cameras, and spectrographs. Ordinary glass lenses and prisms are insufficiently transmissive at wavelengths longer than 2 μm to be useful in an infrared optical system. Instead, infrared transmissive materials such as calcium fluoride, zinc selenide, silicon, and germanium can be fabricated into lenses and prisms. Calcium fluoride, for example, is transparent to infrared light of wavelengths as long as 10 μm. Mirrors and diffraction gratings in infrared optical systems are often gold-coated since gold is a particularly efficient reflector across the entire infrared region of the spectrum. Many instruments depend on the use of wavelength-band-limiting filters for their operation. These filters are produced by coating an infrared transmissive substrate with multiple alternating layers of material of differing refractive index. With proper tuning of the thickness and number of layers, the constructive and destructive interference of the electromagnetic waves reflected from the layer boundaries can efficiently transmit a well-defined range of wavelengths. *See* INTERFERENCE FILTERS; OPTICAL MATERIALS; REFLECTION OF ELECTROMAGNETIC RADIATION.

Astronomical targets. The targets of infrared observations include ordinary stars and galaxies, planets, brown dwarfs, young stellar objects, evolved stars, starburst galaxies, and redshifted radiation.

Ordinary stars and galaxies. Although popular interest in infrared astronomy focuses on exotic objects observable only using infrared light, infrared observations continue to play a fundamental role in understanding the more pedestrian stars and galaxies that constitute most of the visible-wavelength universe. Just as at visible wavelengths, where the colors and luminosities of stars divide them into distinct spectral classes via the Hertzsprung-Russell diagram, stars can be classified as to their temperature and luminosity class via their infrared "colors." In this case, color refers to the ratio of the brightness of an object compared at two different infrared wavelengths. Sources with proportionally more flux at longer infrared wavelengths are referred to as "red" by infrared astronomers in analogy to objects which literally appear red using visible light. Color-based distinctions can also be made among classes of galaxies. At infrared wavelengths, interstellar dust within the Milky Way Galaxy becomes only a minor hindrance to seeing stars across or galaxies beyond the galactic plane, permitting a more complete census of these objects. At infrared wavelengths longer than 10 μm, infrared light emitted from dust warmed by stars in normal galaxies begins to dominate the infrared light from stars, making dusty spiral galaxies prime targets for mid-infrared observations. *See* GALAXY, EXTERNAL; HERTZSPRUNG-RUSSELL DIAGRAM; MILKY WAY GALAXY; STAR.

Planets. At the shortest infrared wavelengths, planets and asteroids are observed via reflected sunlight. Their infrared spectra are diagnostic of their surface mineralogy and atmospheric chemistry. The outer gas giants have near-infrared reflected spectra which are dominated by the absorption bands of molecular hydrogen and methane in their atmospheres. At infrared wavelengths longer than 3 μm, thermal radiation from the planets begins to dominate over reflected sunlight. Jupiter, still cooling from its initial formation, emits twice as much energy as it receives from the Sun—nearly all of it at infrared wavelengths. *See* ASTEROID; JUPITER; PLANET; PLANETARY PHYSICS.

Brown dwarfs. An object with a mass less than 8% that of the Sun (equivalent to 80 times the mass of Jupiter) is incapable of burning hydrogen in its core to sustain its luminosity—the hallmark of being a star. At the time of their formation, the interiors of such substellar objects are warmed by their gravitational contraction from an initially diffuse interstellar cloud of gas. Immediately following their formation, their surface temperature mimics those of the coolest stars (2000 K or 3100°F). Even at this most luminous point in their evolution, most of the light emerges in the infrared part of the spectrum. Since they have no internal source of energy, the radiation emerging from their surface causes them to cool over time. An object with 3% of the mass of the Sun cools to a surface temperature of 300 K (80°F) after only about 10^9 years. Infrared surveys of the entire sky are beginning to reveal large numbers of these brown dwarfs, and they appear to be more common than ordinary stars in the Milky Way Galaxy.

Young stellar objects. Young stellar objects (YSOs) are newly formed or forming stars. These stars are being assembled by gravity out of dense (1000 atoms per cubic centimeter) interstellar clouds. The environment surrounding the forming star is naturally very dusty, and the slight rotation of the natal cloud combined with gravity drives the infalling material to form a thin flattened disk around the star. This disk may be heated either by the friction of the infalling material or by radiation from the central star. In either case, the disk temperatures range from 2000 K (3100°F) close to the star to 20 K (−424°F) in the outer region of the disk. Virtually all of the disk emission emerges at infrared wavelengths. Planets accrete within these disks, and infrared observations provide the primary astrophysical insight into the process of planetary formation. *See* PROTOSTAR.

Evolved stars. After exhausting the initial supply of hydrogen in the stellar core, stars reconfigure themselves to liberate the energy of hydrogen shell burning around the core and helium burning within the core. Stellar physics dictates that the star must grow to large size and become cool at its surface in order to dissipate the energy being produced within. At this stage, some fraction of stars become unstable to pulsation. Their outer envelopes are weakly bound and easily lost. Dust grains condensing in the expanding envelope can completely enshroud the star. Under these circumstances, the observed radiation emerges entirely in the infrared part of the spectrum as the dense surrounding dust absorbs the short-wavelength starlight and, warmed only to a temperature of a few hundred kelvins, emits exclusively infrared radiation. Stars slightly more massive than the Sun, which are common throughout the Milky Way Galaxy, evolve through this stage. Since these stars are intrinsically luminous and numerous, and because the emergent infrared luminosity penetrates the dust which permeates the Milky Way Galaxy, these stars are ideal tracers of the structure of the Galaxy. *See* GIANT STAR; STELLAR EVOLUTION.

Starburst galaxies. Although normal galaxies emit most of their light at the boundary between the visible and infrared portions of the spectrum, galaxies undergoing active bursts of star formation can produce most of their radiation at wavelengths of 10 μm or longer. The radiation emerges in this part of the spectrum because the star-forming regions are embedded in dust clouds which absorb the starlight and, having been warmed to temperatures of tens of kelvins, reradiate energy largely in the infrared portion of the spectrum. *See* STARBURST GALAXY.

Ultraluminous infrared galaxies. The gravitational interaction between two gas-rich galaxies can induce both objects to undergo an extensive burst of star formation. The resulting energy release can augment the flux of the galaxy by a factor of 10 or more with most of the radiation arising in the infrared part of the spectrum. Such "ultraluminous" infrared galaxies are among the most luminous galaxies in the universe.

Redshifted radiation. The apparent Doppler shift due to the expansion of the universe causes the ultraviolet and visible light originally emitted by extremely distant stars and galaxies to be shifted into the infrared part of the spectrum. Redshifts this large originate from objects at distances of 10^{10} light-years or more. Since the light collected from these objects was emitted by them 10^{10} years ago, these observations probe the state of the universe at the earliest times. Many observational investigations of the early universe focus on collecting and analyzing infrared light from the most distant objects known. Bolometer arrays capable of imaging at a wavelength of 300 μm are now able to detect the first galaxies which formed following the big bang. *See* BIG BANG THEORY; COSMOLOGY; REDSHIFT.

Ground-based infrared astronomy. Telescopes dedicated to visible-wavelength astronomy are also effective collectors of infrared radiation. Equipped with infrared focal-plane-array imagers and spectrographs, these telescopes can observe the infrared universe through the accessible atmospheric windows (Fig. 1). Such observations are particularly effective shortward of 2 μm, where thermal emission from the telescope is negligible relative to other backgrounds. Nearly every telescope larger than 2 m (80 in.) in aperture has one or more dedicated infrared instruments that share time with visible-wavelength images and spectrographs. Since moonlight interferes strongly with visible observations but hardly affects infrared observations, instruments are scheduled in synchrony with the phases of the Moon. Near full moon, most large telescopes conduct infrared observations.

Ground-based infrared observations are less susceptible to blurring due to atmospheric turbulence than visible-wavelength observations. Under ideal circumstances, only diffraction of light through the telescope aperture should blur stellar images. At visible wavelengths, atmospheric turbulence broadens images by a factor of 50 or more compared with the diffraction limit. Modern computers have enabled the incorporation of flexible mirrors into telescope optical paths which compensate for the blurring effects of the atmosphere. At visible wavelengths, the atmosphere changes too rapidly to permit optimal removal of the effects of atmospheric turbulence. The changes occur more slowly at infrared wavelengths, and such adaptive-optics systems can compensate for the atmosphere's effects and yield images with diameters limited by diffraction alone (**Fig. 3**). In addition to providing images of finer detail, the sharper images concentrate more light into a smaller region, making it possible to detect fainter objects. *See* ADAPTIVE OPTICS.

Infrared space missions. The opacity of the Earth's atmosphere at most infrared wavelengths and the need to place a telescope in an environment where the entire telescope structure can be cooled to cryogenic temperatures have motivated a number of extremely successful satellite missions largely devoted to infrared astronomy.

IRAS. The *Infrared Astronomy Satellite* (*IRAS*) conducted the first survey of nearly the entire sky at wavelengths of 12, 25, 60, and 100 μm. The

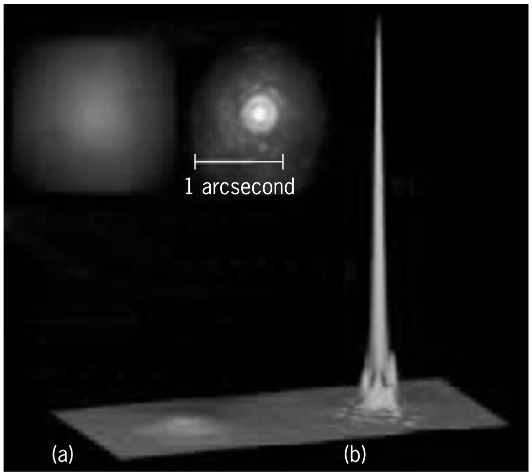


Fig. 3. Example of the improvement in both sharpness and intensity of a stellar image with the application of adaptive optics to compensate for atmospheric turbulence. Plots of image intensity with adaptive optics (*a*) turned off and (*b*) on show improvement both in resolution from 0.6 to 0.068 arcsecond (as measured by the full width of the peak at half its maximum amplitude) and in peak intensity, enabling the detecction of fainter objects. Inset shows the corresponding images. Images were obtained at the Canada-France-Hawaii Telescope (aperture of 3.3 m or 130 in.) at a wavelength of 0.936 μm with an exposure time of 30 s. (*University of Hawaii*)

three longest wavelengths are accessible almost exclusively from space. Launched by the National Aeronautics and Space Administration in 1983, *IRAS* contained a 0.6-m (24-in.) telescope cooled to liquid helium temperature and a focal plane of dozens of individual infrared detectors which scanned across the sky. The resulting catalogs of 350,000 stars and galaxies, along with accompanying images of most of the sky, remain a fundamental scientific resource. As might be expected when opening an entirely new wavelength regime to sensitive observation, the *IRAS* mission revealed a number of fundamentally new phenomena, including residual dust disks around main-sequence stars (presumably the debris of evolving solar systems), ultraluminous infrared galaxies, and tenuous filaments of interstellar dust now known as infrared cirrus.

ISO. The *IRAS* mission placed the infrared sky in context with an all-sky survey. In 1995, the European Space Agency launched the *Infrared Space Observatory* (*ISO*), the first infrared observatory with the capability of making detailed photometric and spectroscopic studies of individual objects at wavelengths ranging from 2.5 to 240 μm. The *ISO*'s 0.6-m (24-in.) liquid-helium-cooled telescope operated for $2^{1}/_{2}$ years, making more than 20,000 independent observations of astronomical sources. The telescope's focal plane included an infrared camera (ISOCAM), with sensitivity from 2.5 to 17 μm, which imaged the sky onto two different 32 × 32-element infrared focal-plane arrays; an infrared photometer and polarimeter (ISOPHOT), which used individual detectors to measure precisely the quantity of infrared radiation from celestial sources with detectors sensitive between 2.5 and 240 μm; and a pair of spectrometers observing at wavelengths between 2.4 and 200 μm. *ISO* observations have contributed to nearly every area of astrophysics, ranging from spectroscopic studies of planetary atmospheres to infrared imaging of forming galaxies in the early universe.

SIRTF. Scheduled for launch into an Earth-trailing orbit in 2001, the *Space Infrared Telescope Facility* (*SIRTF*) will place a sophisticated battery of infrared cameras and spectrographs behind a 0.85-m (33.5-in.) liquid-helium-cooled telescope with a 5-year cryogen lifetime. *SIRTF*'s infrared array camera (IRAC) will contain four 256 × 256-element infrared arrays spanning the wavelength range from 3.6 to 8.0 μm. These arrays are substantially larger and more sensitive than those available on *ISO* and illustrate the rate of advance of infrared imaging technology. *SIRTF* will have an imaging capability to wavelengths as long as 160 μm and spectroscopic coverage between 4 and 40 μm. *SIRTF* joins the *Compton Gamma-Ray Observatory*, the *Chandra X-ray Observatory*, and the Hubble Space Telescope (observing at visible and near-infrared wavelengths) as one of NASA's "great observatories," which operate across the electromagnetic spectrum. *SIRTF* will illuminate nearly every area of astrophysics and is particularly well suited to revealing the abundance of cool brown dwarfs and ultraluminous infrared galaxies.

Prospects. More ambitious plans for space telescopes emphasize infrared observational capability because of the range of astrophysics addressed in this wavelength regime as well as relaxed design constraints on telescopes, since infrared wavelengths are several times larger than visible-light wavelengths. NASA's long-term goal involves detecting and imaging planets around other stars. Infrared imaging and interferometry play a central role in the realization of this objective. Since the planets of interest naturally reside next to an extremely bright star, an interferometer—which combines the light from two or more well-separated telescopes—is required to isolate the planet's feeble light from that of the star.

Interferometry has been pioneered largely at infrared wavelengths from the ground, since the Earth's atmosphere appears more stable when observing infrared light than when observing visible light. In addition, the contrast between the star and adjacent planet is substantially improved at infrared wavelengths, making the planet easier to discern in the star's glare. Infrared spectroscopic observations are diagnostic of water and ozone in planetary atmospheres and will be fundamental in assessing the probability that life could arise or even has developed on extrasolar planets.

Michael F. Skrutskie

Bibliography. G. Basri, The discovery of brown dwarfs, *Sci. Amer.*, 282(4):76–83, April 2000; H. Freudenreich, Deconstructing the Milky Way, *Amer. Sci.*, 87(5):418–427, September–October 1999; C. Kraan-Korteweg and O. Lahav, Galaxies behind the Milky Way, *Sci. Amer.*, 279(4):50–57, October 1998; M. Skrutskie, Charting the infrared sky, *Sky Telesc.*, 94(2):46–47, August 1997.

Infrared imaging devices

Devices that convert an invisible infrared image into a visible image. Infrared radiation is usually considered to span the wavelengths from about 0.8 or 0.9 micrometer to several hundred micrometers; however, most infrared imaging devices are designed to operate within broad wavelength regions of atmospheric transparency, that is, the atmospheric windows. At sea level, for horizontal paths of a few kilometers' length, these are approximately at 8–14 μm, 3–5 μm, 2–2.5 μm, 1.5–1.9 μm, and wavelengths shorter than 1.4 μm. The radiation available for imaging may be emitted from objects in the scene of interest (usually at the longer wavelengths called thermal radiation) or reflected. Reflected radiation may be dominated by sunlight or may be from controlled sources such as lasers used specifically as illuminators for the imaging device. The latter systems are called active, while those relying largely on emitted radiation are called passive. Active optical imaging systems were developed to achieve a nighttime aerial photographic capability, and work during World War II pushed such systems into the near-infrared spectral region. Development of passive infrared imaging systems came after the war, but only the advent of lasers allowed creation of active infrared imagers at wavelengths much longer than those of the photographic region. Striking advances have been made in active infrared systems which utilize the coherence available from lasers, and hybrid active-passive systems have been studied intensively. *See* INFRARED RADIATION; LASER.

Although developed largely for military purposes, infrared imaging devices have been valuable in industrial, commercial, and scientific applications. These range from nondestructive testing and quality control to earth resources surveys, pollution monitoring, and energy conservation. Infrared images from aerial platforms are used to accomplish "heat surveys," locating points of excessive heat loss. An example is shown in **Fig. 1***a*. As discussed below, calibration allows association of photographic tones in this figure with values of apparent (that is, equivalent blackbody) temperatures. Dark areas in the figure are "colder" than light ones. *See* NONDESTRUCTIVE EVALUATION; REMOTE SENSING.

Scanning systems. Infrared imaging devices may be realized by electrooptical or optomechanical scanning systems. All have an optical head for receiving the infrared radiation and a display for the visible image. These are connected by electronics for the passage of electrical signals from the detector element(s) to the display input. Signal processing may be incorporated in the electronics to selectively enhance or reduce features in the produced visible image. For example, in Fig. 1*b* a "level-slice" technique presents in white all areas (mainly rooftops) with apparent temperatures between −7.9 and −8.9°C or 17.4 and 15.6°F. (The ambient air temperature was −5°C or 23°F.) Black regions in the figure correspond to apparent temperatures below or above the narrow "sliced" temperature range of the white regions.

Fig. 1. Thermal imagery in the wavelength range 10.4–12.5 μm obtained during flights over Ypsilanti, Michigan, at 2400 hours, November 23, 1975, by the Airborne Multispectral Scanner operated by the Environmental Research Institute of Michigan. (*a*) Calibrated thermal imagery. (*b*) Signal-processed thermal imagery of same scene. (*From F. Tannis, R. Sampson, and T. Wagner, Thermal imagery for evaluation of construction and insulation conditions of small buildings, Environmental Research Institute of Michigan, ERIM Rep. 116600-12-F, July 1976*)

Optomechanical methods such as rotating prisms or oscillating mirrors may be used to sample or scan the spatial distribution of infrared radiation in either the object or image plane. Electrooptical imaging devices may use an electron beam (for example, vidicons) or charge transport in solids (for example, charge-coupled devices, or CCDs) to scan the infrared image formed by the optics of the device. This image-plane scanning places more stringent requirements upon the optics for image quality off-axis than does use of mechanically moved optical elements placed before the entrance aperture of the system. Intensive development of pyroelectric vidicons, detector arrays, and infrared charge-coupled devices (IRCCDs) has taken place, reflecting the critical role played by the detector element in all infrared systems. The spectral, spatial, and temporal responses of detectors are the major factors in determining the wavelength regions, the spatial resolution, and the frequency response (that is, the time constant) of imaging devices. *See* CHARGE-COUPLED DEVICES.

Detector arrays. Optomechanical scanning methods often stress the response time of the detector-electronics package. As a result, multiple detectors or detector arrays are sometimes incorporated in the focal planes, resulting in partially electronically scanned optomechanical systems. The technology for use of a linear array of detector elements (often lead selenide and indium antimonide detectors for the 3–5-μm region, and mercury-doped germanium and mercury cadmium telluride for the 8–14-μm window) is well developed, and the use of a two-dimensional array or matrix of detectors has been studied. Optomechanical imagers incorporating such arrays allow the use of time delay and integration (TDI) of the signals to improve the resulting signal-to-noise ratios.

Solid-state components such as charge-coupled devices afford the opportunity for implementation of signal processing directly at the focal plane. Two approaches have been undertaken to attain the focal-plane array technology of infrared charge-coupled devices. In one, the development of a hybrid device, an infrared detector matrix of any suitable photodetector material, for example, indium antimonide, mercury cadmium telluride, and lead tin telluride, is mated with a conventional silicon charge-coupled device. Thus two solid-state wafers or "chips" are integrated to obtain an infrared charge-coupled device. In the other, the goal is a monolithic chip, one incorporating the photodetection, charge generation, and charge transfer in a structure made of essentially one material. Candidate materials include impurity-doped silicon, indium antimonide, and mercury cadmium telluride. The hybrid device technology can be implemented more readily than that needed for monolithic infrared charge-coupled devices. The development of infrared charge-coupled devices with the number of detecting elements in a sufficiently closely packed array required for high-performance infrared imaging devices involves additional difficulties.

Scanning motion. Some optomechanical imagers produce a two-dimensional scan entirely by movement of components of the device itself; others utilize the motion of a platform such as an aircraft or satellite. The first kind of system includes the forward-looking infrared (FLIR) or framing imagers which usually scan in televisionlike raster patterns and display, synchronously if done in real time, a visible image corresponding to the spatial distribution of infrared radiation. These visible image outputs have been named thermographs and thermograms. Commercially available imaging devices of this type have used horizontally and vertically rotating germanium prisms, mirrors oscillated in orthogonal directions, two mirrors and a six-sided rotating prism, and other schemes to produce images at relatively slow rates from 16 per second to less than a quarter of a frame per second. Higher-performance systems have been produced for military purposes. The second class of imaging systems includes those often called line scanners or thermal mappers. One such system, the 12-channel Airborne Multispectral Scanner operated by the Environmental Research Institute of Michigan (ERIM; **Fig. 2**), includes two thermal radiation channels, at 8.2–9.3 μm and 10.4–12.5 μm whose magnetic-tape recorder output was processed to produce the thermal imagery in Fig. 1.

Multispectral imaging systems can be used to collect large amounts of data that may be processed to create special-purpose images. Advanced image data processing techniques are being used in remote sensing applications such as mineral detection and assessment, vegetation cover classification, agricultural crop monitoring and management, and urban analyses and planning. If the system has more than about 30 spectral channels with relatively fine spectral resolution, imaging spectroscopy can be carried out. These hyperspectral imaging systems, compared with multispectral ones, have data outputs better described by more complex representations than simple images. Hypercubes, many-dimensional vectors that need multidimensional data analysis methodologies, have become especially useful. Industrial and laboratory applications now include spatially resolved infrared microspectroscopy. Such systems are becoming available as a direct result of the creation of modern focal-plane arrays (FPAs) for use in, for example, fast Fourier transform infrared (FT-IR) imaging systems. A simple illustration of a system for remote sensing applications from airborne or space-based platforms would look similar

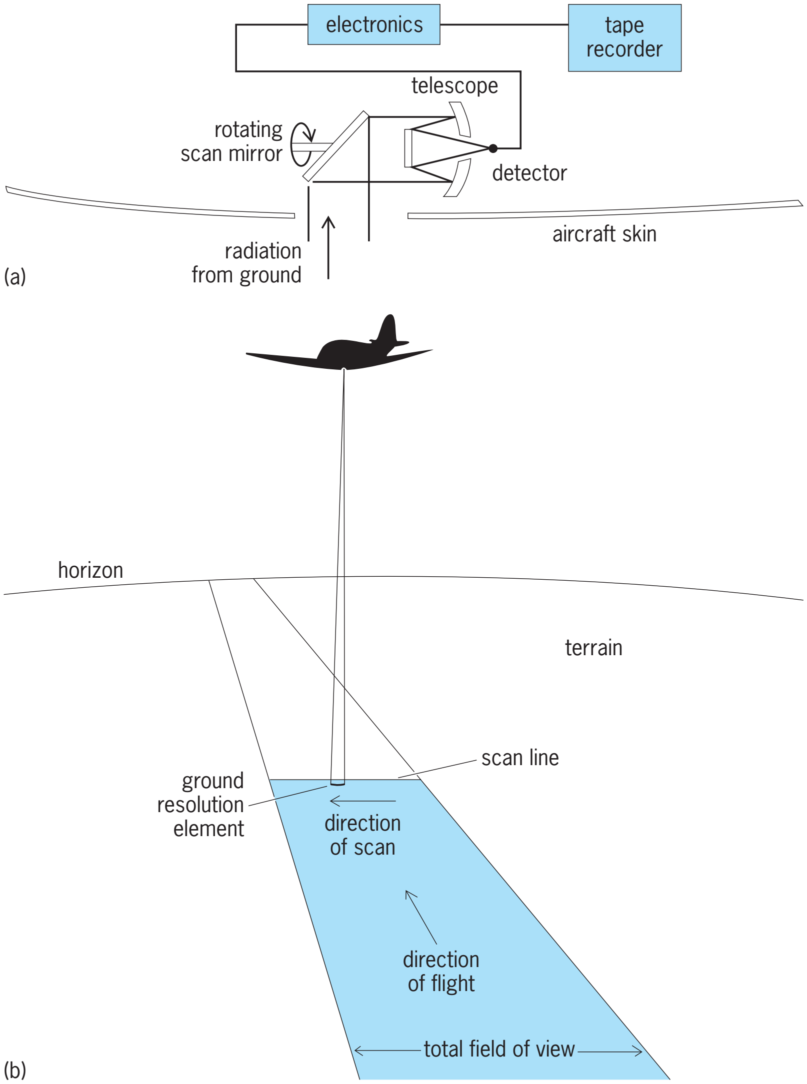


Fig. 2. Airborne Multispectral Scanner. (*a*) Schematic diagram of equipment. (*b*) Scanning operation, utilizing motion of aircraft. (*After F. Tannis, R. Sampson, and T. Wagner, Thermal imagery for evaluation of construction and insulation conditions of small buildings, Environmental Research Institute of Michigan, ERIM Rep. 116600-12-F, July 1976*)

to Fig. 2*a*, with the detector replaced by an entrance aperture for a multispectral or hyperspectral dispersing element, such as a prism or grating or interferometric subsystem. *See* FOURIER SERIES AND TRANSFORMS.

Characterization of output. The instantaneous field of view (IFOV) or resolution element of imaging systems is always geometrically filled by the radiating source, so that the output of the device is a response to changes in amount of radiation from field of view to field of view. These changes are best characterized in terms of radiance L, the radiant flux per unit area per unit solid angle, usually in a selected spectral band of wavelengths, $\Delta\lambda$. Even in the infrared regions, the radiance variation may be ascribed to changes in reflectance, incident radiation, emissivity, or temperature. By restriction to wavelengths dominated by emission, the so-called thermal wavelengths longer than 3.5 μm, the radiance change can be described by the equation below, where T is the

$$\Delta L = \frac{\partial L}{\partial T}\,\Delta T + \frac{\partial L}{\partial \epsilon}\,\Delta\epsilon$$

absolute temperature and ϵ is the emissivity. Contributions due to $\Delta\epsilon$ are usually treated as changes in an equivalent blackbody temperature by setting $\epsilon = 1$ and $\Delta\epsilon = 0$. Then T represents an equivalent blackbody temperature, and the radiance variation can be ascribed entirely to a value of ΔT. That value of ΔT corresponding to a radiance difference which will just produce a signal-to-noise ratio of 1 is called the noise equivalent temperature difference (NETD). One can also characterize the performance of an imaging system by a noise equivalent emissivity difference or even a noise equivalent reflectivity difference. The use of noise equivalent temperature difference as a figure of merit for thermal imagers is obviously more appropriate. For the higher-performance forward-looking infrared systems, a useful figure of merit is the minimum resolvable temperature difference (MRTD), a figure of merit which includes the characteristics of the display and of the observer as well. *See* EMISSIVITY; HEAT RADIATION.

Display. If a visible image is the desired final output of an infrared imaging device, it may be displayed as a conventional television picture by means of a cathode-ray tube (CRT). Cathode-ray-tube technology has been developed to a level that is suitable, and research has been undertaken toward creation of satisfactory flat panel displays using liquid crystal elements, light-emitting diodes, or plasma panels. Systems not requiring a real-time image display may utilize analog or digital data storage or transmission systems, which then are used to produce permanent visual records such as photographs. High-resolution "hard copy" images can be produced by sophisticated systems using electron-beam or laser recording on film. Complex signal-processing techniques are easily introduced before the final image recording is made. *See* CATHODE-RAY TUBE; LIGHT-EMITTING DIODE; LIQUID CRYSTALS.

George J. Zissis

Bibliography. J. S. Accetta and D. L. Shumaker (eds.), *The Infrared and Electro-Optical Systems Handbook*, vols. 1–8, ERIM and SPIE, 1993; M. Bass with E. Stryland, D. R. Williams, and W. L. Wolfe (eds.), *Handbook of Optics*, 2d ed., vols. 1 and 2, Optical Society of America/McGraw-Hill, 1995; G. C. Holst, *CCD Arrays, Cameras, and Displays*, 2d ed., SPIE and JCD, 1998; G. C. Holst, *Testing and Evaluation of Infrared Imaging Systems*, 2d ed., SPIE and JCD, 1998; W. L. Wolfe, SPIE Tutorial Texts: *Introduction to Infrared System Design*, 1996, *Introduction to Imaging Spectrometers*, 1997, *Introduction to Radiometry*, 1998, and with G. J. Zissis (eds.), *The Infrared Handbook*, 1978, SPIE/ERIM-IRIA.

Infrared lamp

A special type of incandescent lamp that is designed to produce energy in the infrared portion of the electromagnetic spectrum. The lamps produce radiant thermal energy which can be used to heat objects that intercept the radiation. All incandescent lamps have radiation in three regions of the electromagnetic spectrum, the infrared, the visible, and the ultraviolet. An infrared lamp with a filament operating at 2500 K will release about 85% of its energy in the form of thermal radiant energy, about 15% as visible light, and a tiny fraction of a percent as ultraviolet energy. The amount of infrared radiation produced by a lamp is a function of the input wattage of the lamp and the temperature of its tungsten filament. For most infrared lamps, the thermal radiation will be about 65–70% of the input wattage. A typical 250-W infrared lamp radiates about 175 W of thermal energy.

Temperature and energy spectrum. The temperature of the incandescent filament determines the portion of the energy output allocated to each spectral region. As the filament temperature increases, the wavelength at which peak energy output occurs shifts toward the visible region of the spectrum (**Fig. 1**). Lamp filament temperatures are always stated in kelvins (K). *See* HEAT RADIATION; TEMPERATURE.

Types. The lamps are supplied in two shapes (**Fig. 2**). The most common shape, for general use, is the R lamp (Fig. 2*a*), since the reflector unit is built in and the lamp needs only a suitable socket to form an infrared heating system. These lamps are available in 125-, 250-, and 375-W ratings, all in glass bulbs with a nominal reflector diameter of 12.5 cm (5.0 in.) and voltage rating of 115–125 V.

The other type of infrared lamp, the tubular quartz bulb lamp (Fig. 2*b*), is used with a separate external reflector designed to distribute the heat as desired. More than 30 sizes of the tubular quartz bulb lamps are available in 375–5000-W ratings. Lighted lengths range from 12.5 cm (5 in.) to 127 cm (50 in.), overall lengths (including connecting terminals) from 22.38 cm (8.8 in.) to 138.68 cm (53.8 in.), and tube diameter is 0.95 cm (0.375 in.). Voltage ratings are a

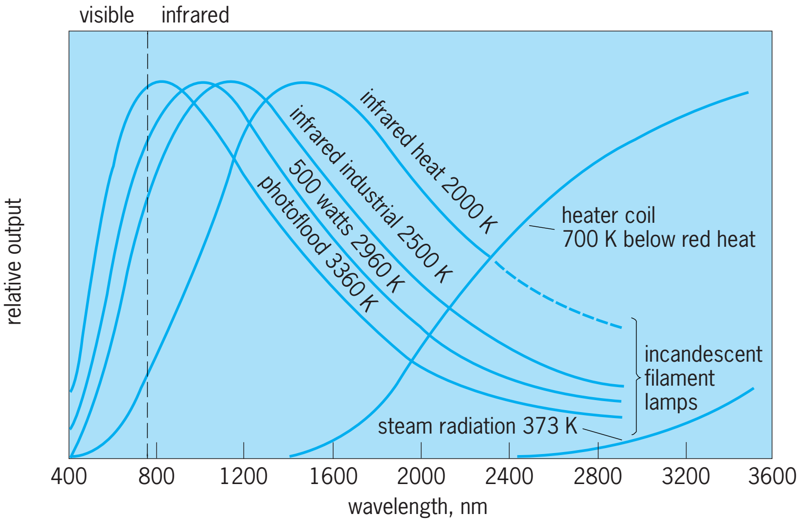
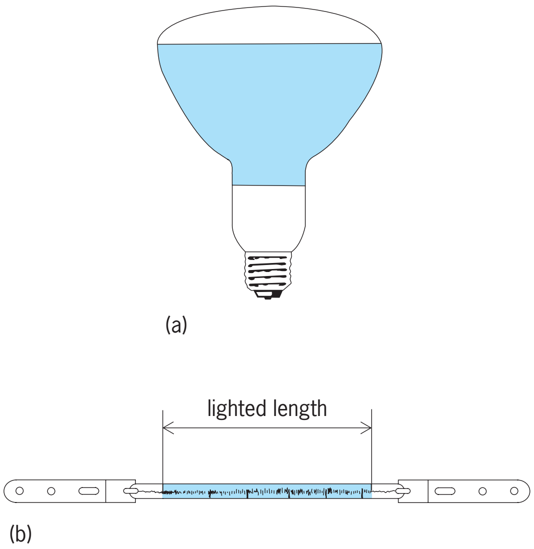


Fig. 1. Relative energy output of four incandescent lamps as a function of wavelength and filament temperature. As temperature increases, more of the lamp's energy shifts to shorter wavelengths. °F = (K × 1.8) − 460. (*After Illuminating Engineering Society, IES Lighting Handbook, 8th ed., 1993*)



(a)

(b)

Fig. 2. Typical shapes of infrared lamps. (*a*) R lamp, with built-in reflector unit. (*b*) Tubular configuration with a quartz bulb.

function of length, with nominal values ranging from 120 V for short tubes to 960 V for the longest tubes.

Because the lamp bases become quite hot, R lamps (Fig. 2*a*) should be used only in porcelain sockets, rather than in the common brass shell socket which would fail with prolonged use, posing a fire hazard. The R lamp is available with the glass face of the bulb stained red or ruby, reducing the amount of light emitted by the lamp. This reduces the undesirable glare in situations where the lamps must be in sight of the workers.

Uses. The major advantage of infrared heating is that it is possible to heat a surface that intercepts the radiation without heating the air or other objects that surround the surface being heated. Infrared lamps have many uses, some of which will be discussed.

Paint drying. Many industrial infrared ovens are used to dry painted surfaces as they pass by on a conveyor belt. Drying times are much shorter than for other methods of drying paint.

Evaporative drying. Infrared lamps can be used to remove moisture or solvents from the surface of objects by speeding the evaporation process. Porous materials with internal moisture require longer exposure, since the entire object must be heated by conducting heat from the surface to the interior.

Farm heating of animals. The lamps can be used to heat brooders and to keep other livestock warm without having to heat a large, poorly insulated shed.

Food. Many R lamps are used in restaurant kitchens to keep food warm while it is waiting to be served to customers.

Comfort heating. Areas under theater marquees can be made more pleasant for patrons waiting in line to buy tickets, and sidewalk cafes and similar areas can be made comfortable in relatively chill weather by installing suitable lamp systems.

Other uses. These include therapeutic heating of portions of the body to relieve muscle strains, tobacco drying, textile and wallpaper drying, and drying of printing inks so that press operation can be speeded up without causing image offset. *See* THERMOTHERAPY.

Design of heating systems. Most infrared heating ovens or systems are designed by calculating the number and wattage of lamps, plus experimenting with the system after it is constructed. The latter is necessary because the number of variables is large, and precise data on the magnitude of the variables are often difficult to obtain.

Voltage and lifetime. Moderate variations in voltage will alter lamp operation somewhat. A 1% increase in voltage will increase both the input and radiated power in watts by 1.5–2%, and it will also increase the filament temperature and the light from the lamp. Most lamps are rated for use over a range of voltage, such as 115–125 V for a lamp intended for use on a 120-V power system. The life of all infrared lamps exceeds 5000 h. The greatest cause of life reduction is subjecting the lamp to vibration.

Exposure to radiation. Exposure of persons to moderate levels of infrared is not known to be harmful. High levels could cause burns, but the exposed individual feels the heat and would normally move away before a burn actually occurred. *See* INCANDESCENT LAMP; INFRARED RADIATION.

G. R. Peirce; Robert Leroy Smith

Bibliography. K. J. Button (ed.), *Infrared and Millimeter Waves*: *Sources of Radiation*, vol. 1, 1979; J. D. Hall, *Industrial Applications of Infrared*, 1947; Illuminating Engineering Society, *IES Lighting Handbook*, 8th ed., 1993; M. La Toison, *Infrared and Its Thermal Applications*, 1964; W. L. Wolfe and G. J. Zissis (eds.), *The Infrared Handbook*, 1978, reprint 1985.

Infrared radiation

Electromagnetic radiation in which wavelengths lie in the range from about 1 micrometer to 1 millimeter. This radiation therefore has wavelengths just a little longer than those of visible light and cannot be seen with the unaided eye. The radiation was discovered in 1800 by William Herschel, who used a prism to refract the light of the Sun onto mercury-in-glass thermometers placed just past the red end of the visible spectrum generated by the prism. Because the techniques and materials used to collect, focus, detect, and display infrared radiation are different from those of the visible, and because many of the applications of infrared radiation are also quite different, a technology has arisen, and many scientists and engineers have specialized in its application. *See* ELECTROMAGNETIC RADIATION.

Infrared techniques. A complete infrared system consists of a radiating source and background, intervening medium, optical system, detector, electronics, and display.

Sources. The source can be described by the spectral distribution of power emitted by an ideal radiating body. This distribution is characteristic of the temperature of the body. A real body is related to it by a radiation efficiency factor also called emissivity. It is the ratio at every wavelength of the emission of a real body to that of the ideal under identical conditions. **Figure 1** shows curves for these ideal bodies radiating at a number of different temperatures. The higher the temperature, the greater the total amount of radiation. The total number of watts per square meter is given by $5.67 \times 10^{-8}T^4$, where T is the absolute temperature in kelvins (K). Higher temperatures also provide more radiation at shorter wavelengths. This is evidenced by the maxima of these curves moving to shorter wavelengths with higher temperatures. Ideal radiators are also called blackbodies. *See* BLACKBODY; EMISSIVITY; HEAT RADIATION.

The sources can be either cooperative or uncooperative. Some examples of the former include tungsten bulbs (sometimes with special envelopes), globars, and Nernst glowers. These are made of rare-earth oxides and carbons. They closely approximate blackbodies and are used mostly for spectroscopy. Lasers have been used in special applications. Although they provide very intense, monochromatic, and coherent radiation, they are limited in their spectral coverage. The principal infrared lasers have been carbon dioxide (CO_2) and carbon monoxide (CO) gas lasers and lead-tin-tellurium (PbSnTe) diodes. *See* INFRARED LAMP; LASER.

Transmitting medium. The radiation of one of these sources propagates from the source to the optical collector. This path may be through the vacuum of outer space, 3 ft (1 m) of laboratory air, or some arbitrary path through the atmosphere. **Figure 2** shows a low-resolution transmission spectrum of the atmosphere and the transmissions of the different atmospheric constituents. Two of the main features are the broad, high transmission regions between 3 and 5 μm and between 8 and 12 μm. These are the spectral regions chosen for most terrestrial applications. **Figure 3** shows a small section of this spectrum, illustrating its complexity. The radiation from the source is filtered by the atmosphere in its propagation so that the flux on the optical system is related to both the source spectrum and the transmission spectrum of the atmosphere.

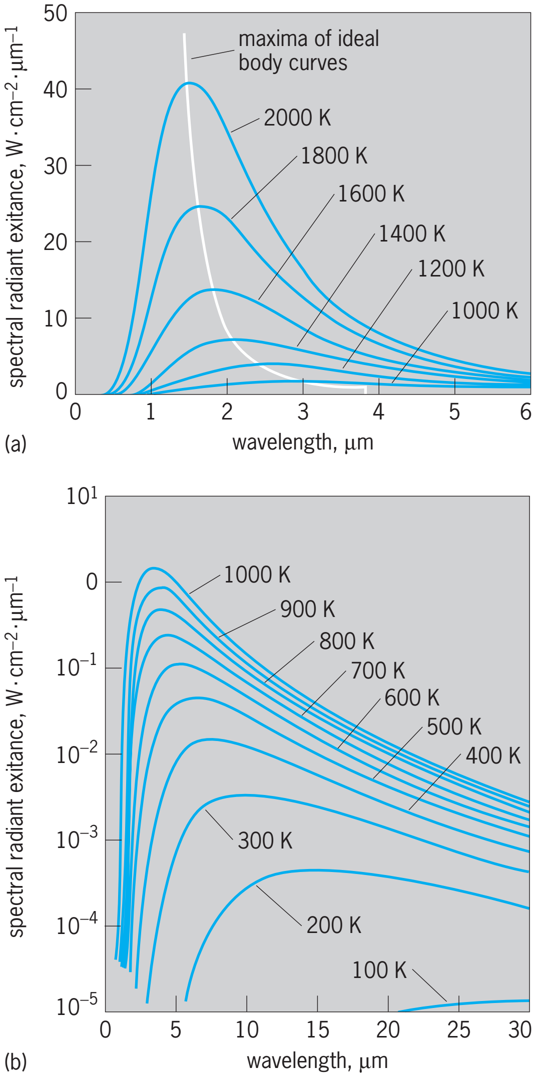


Fig. 1. Radiation from ideal bodies (blackbodies) at different temperatures, shown on (*a*) a linear scale and (*b*) a logarithmic scale. °F = (K × 1.8) − 460.

Optical system. A lens, mirror, or a combination of them is used to focus the radiation onto a detector. Since glass is opaque in any reasonable thickness for radiation of wavelengths longer than 2 μm special materials must be used. The **table** lists the properties of some of the most useful materials for infrared instrumentation. In general, these are not as effective as glass is in the visible, so many infrared optical systems use mirrors instead. The mirrors are characterized by the blank and by its coating. Blanks are usually made of aluminum, beryllium, or special silica materials. The choices are based on high strength, light weight, and good thermal and mechanical stability. They are coated with thin evaporated layers of aluminum, silver, or gold. The reflectivities of thin-film metallic coatings increase with wavelength, and the requirements for surface smoothness are also less stringent with increasing wavelength. *See* MIRROR OPTICS; OPTICAL MATERIALS.

Detectors. Photographic film is not useful for most of the infrared spectrum. The response of the silver halide in the film, even when specially processed, is only sensitive to about 1.2-μm. It cannot respond to radiation in the important 3–5-μm and 8–12 μm atmospheric transmission bands. If there were a film for the infrared, it would have to be kept cold and dark before and after exposure. Accordingly, infrared systems have used point (elemental) detectors or arrays of them. These detectors are based either on the generation of a change in voltage due to a change in the detector temperature resulting from the power focused on it, or on the generation of a change in voltage due to some photon-electron interaction in the detector material. This latter effect is sometimes called the internal photoelectric effect. Electrons which are bound to the different atomic sites in the crystal lattice receive a quantum of photon energy. They are freed from their bound lattice positions and can contribute to current flow. The energy in electronvolts required to do this is $1.24/\lambda$ (where λ is the wavelength in micrometers). Thus only a very small binding energy, about 0.1 eV, is permitted in photon detectors. The thermal agitation of the lattice could cause spontaneous "emission," so most infrared photodetectors are cooled to temperatures from 10 to 100 K (-424 to $-280°$F). This does not affect the speed of response of photodetectors, which depends upon photon-electron interactions, but it does slow down thermal detectors. These are sometimes cooled because they generally respond to a relative change in temperature (and for a lower temperature, a small absolute change gives a larger relative change), and thermal noise is also reduced at low temperatures. *See* BOLOMETER; PHOTOCONDUCTIVE CELL; PHOTODIODE; PHOTOTRANSISTOR; PHOTOVOLTAIC CELL; RADIOMETRY; THERMOCOUPLE.

Electronic circuitry. The voltage or current from the detector or detector array is amplified by special circuitry, usually consisting of metal oxide semiconductor field-effect transistors (MOSFETs) which are designed for low-temperature operations. The amplified signals are then handled very much like television signals. One important feature of most of these systems is that the system does not yield a direct response; only changes are recorded. Thus a "dc restore" or absolute level must be established with a thermal calibration source. The black level of the display can then be chosen by the operator. *See* TRANSISTOR.

Reticle. A reticle or chopper is an important feature of nonimaging systems. A typical reticle and its use are shown in **Fig. 4**. An image of the scene is portrayed on a relatively large fraction of the reticle—anywhere from 10 to 100%. The lens just behind it collects all the light that passes through the reticle and focuses it on a detector. The reticle is rotated. If the scene is uniform, there will be no change in the detector output. However, if the scene has a point source (like the image of the hot exhaust of an aircraft engine), then a point image is formed on the reticle and a periodic detector output is generated. The phase and frequency of this output can be used with properly designed reticles to obtain the angular coordinates of the point source. The reticle pattern can also be used to reduce the modulation obtained from almost uniform scenes which radiate from large areas. Reticles are used in most infrared tracking systems, although other schemes are sometimes employed.

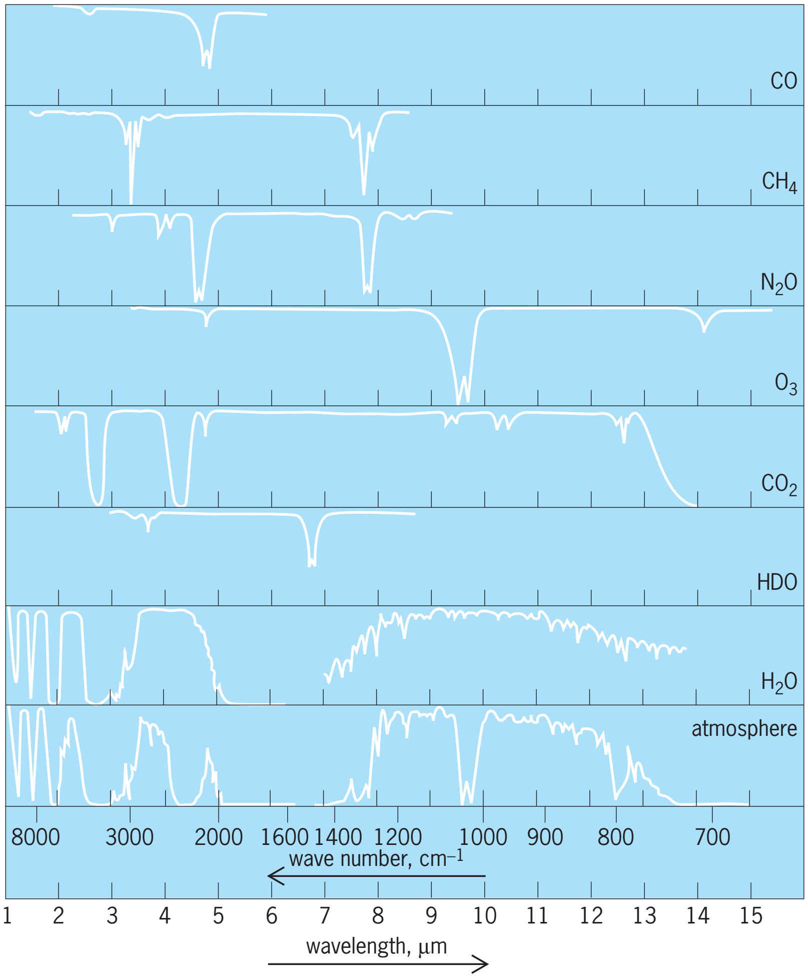


Fig. 2. Low-resolution transmission spectra of the atmosphere and of constituent gases.

Applications. Infrared techniques have been applied in military, medical, industrial, meteorological, ecological, forestry, agricultural, chemical, and other disciplines.

Meteorological applications. Weather satellites use infrared imaging devices to map cloud patterns and provide the imagery seen in many weather reports. Substances at temperatures between 200 and 300 K (-100 and $+80°$F) emit copious amounts of infrared radiation, but are not hot enough to emit in the visible. The Earth has a temperature of approximately 300 K (80°F), and high-altitude clouds are colder (about 250 K or $-10°$F). An infrared sensor placed

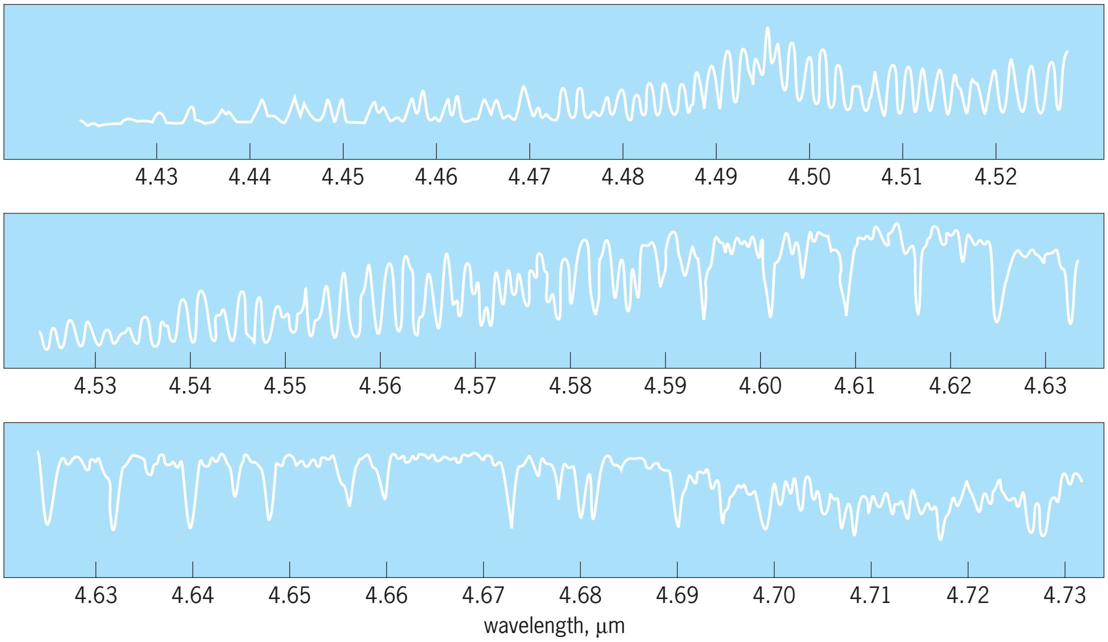


Fig. 3. High-resolution atmospheric transmission spectrum between 4.43 and 4.73 µm.

on a synchronous-orbit satellite can easily sense this difference and present a picture of these temperature patterns. Although the technique is not very widely known, radiometric sensors on lower-altitude satellites can determine the vertical temperature distribution along any path to the Earth. The infrared emission of the Earth's atmosphere is a function of wavelength; a space-borne radiometer senses the emitted radiation at several different wavelengths. The different wavelength bands correspond to greater atmospheric transmission and some "look" deeper into the atmosphere than others. Calculations determine the atmospheric temperature distribution that is required to produce the observed wavelength distribution based on these facts. Although there can be problems with the uniqueness and conversion of such an inversion or fitting process, the results obtained have been accurate to within a few degrees and about a kilometer of altitude. *See* METEOROLOGICAL SATELLITES; REMOTE SENSING.

Medical applications. Infrared imaging devices have also been used for breast cancer screening and other medical diagnostic applications. In most of these applications, the underlying principle is that pathology produces inflammation, and these locations of increased temperature can be found with an infrared imager. The major disadvantage is that localized increases of about 1 K usually need to be detected, and a fair number of nonpathological factors (or at least not the ones in question) can cause equivalent heating. In addition to breast cancer detection (which detects 80% or more cases and has about a 15% false-alarm rate), infrared techniques have been used in the analysis of severe burns, faulty circulation, back problems, and sinus ailments, and has even been proposed to test for medical malingerers.

Airborne infrared imagers. Airborne infrared imagers have been used to locate the edge of burning areas in forest fires. Typically, a forest fire in the western United States is ignited by the lightning of a late afternoon thunderstorm; the valley becomes filled with smoke by the next morning when the crews arrive. The longer wavelengths of the emitted infrared radiation penetrate the smoke better than the visible wavelengths, so the edges of fire are better delineated.

Properties of materials for infrared instrumentation

Material	Transmission region, µm	Approximate refractive index	Comment
Germanium	2–20	4	Opaque when heated
Silicon	2–15	3.5	Opaque when heated
Fused silica	0.3–2.5	2.2	Strong, hard
Zinc selenide	0.7–15	2.4	Expensive
Magnesium fluoride	0.7–14	1.6	Not very strong
Arsenic sulfur glass	0.7–12	2.2	Not always homogeneous
Diamond	0.3–50	1.7	Small sizes only
Salt	0.4–15	1.5	Attacked by moisture

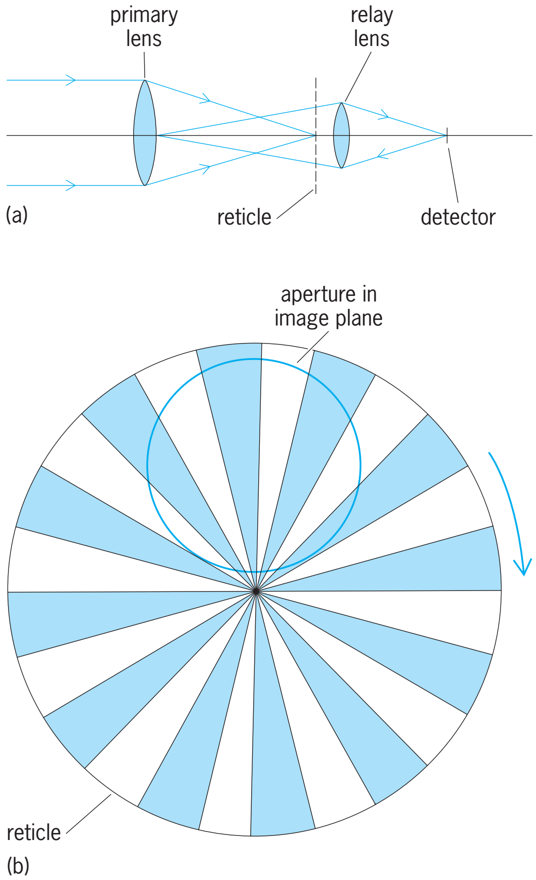


Fig. 4. Reticle system. (*a*) Configuration of components of system. (*b*) Reticle, showing area covered by image.

Thermal pollution contained in the power-plant effluent into rivers has been detected in various locations of the United States, and the viability of crops by assessment of the moisture content has also been accomplished with some degree of success. *See* INFRARED IMAGING DEVICES.

Infrared spectroscopy. Infrared spectrometers have long been a powerful tool in the hands of the analytical chemist. The spectrum of a substance in either absorption or emission provides an unmistakable “fingerprint.” The only problem is to find the spectrum of the substance that matches the unknown. *See* INFRARED SPECTROSCOPY.

Military and space applications. The best-known military techniques are probably the Sidewinder air-to-air missile, which detects the radiation from the exhaust of a jet aircraft, and the spaceborne systems which detect the extremely large amounts of radiation from the plume of launching intercontinental rockets at great distances. Infrared systems have also been used for stabilizing satellites by sensing the Earth’s horizon, for night vision applications, and for perimeter warning and gas detection.

William L. Wolfe

Bibliography. P. Klocek (ed.), *Handbook of Infrared Optical Materials*, 1991; J. L. Miller, *Principles of Infrared Technology*, 1994; M. Schlessinger, *Infrared Technology Fundamentals*, 2d ed., 1994; J. D. Vincent, *Fundamentals of Infrared Detector Operation and Testing*, 1990; W. L. Wolfe and G. J. Zissis, *The Infrared Handbook*, 1978, reprint 1985.

Infrared radiation (biology)

Infrared radiations occupy the span between the visible spectrum and radio waves, and encompass wavelengths 780–400,000 nanometers, neither boundary being precisely delimited. All bodies above absolute zero in temperature are potential sources of infrared radiations; therefore, all organisms are continually exposed to them. About 60% of the Sun’s rays are infrared. Water absorbs infrared radiations strongly, except for the band of transparency between 780 and 1400 nm. Since protoplasm contains much water, it absorbs infrared radiations readily. A large animal absorbs infrared radiations at its surface, only the span from 780 to 1400 nm penetrating as far as the blood vessels. *See* INFRARED RADIATION.

While many substances and even tissues selectively absorb infrared rays, and one might therefore postulate selective effects of these radiations, none have been unequivocally demonstrated except possibly in conjunction with x-rays. The reason is perhaps because quanta of infrared radiation do not excite energy states in molecules other than those excited by conducted heat. *See* X-RAYS.

The essential biological effect of infrared rays depends primarily upon the rise in temperature produced following their absorption, which in turn increases the rate of biological activities in proportion to the temperature change. Because of the prominence of infrared in sunlight, organisms show many adaptations to dissipate or to avoid the heat. The temperature of a large animal or plant may rise temporarily, but the heat is dissipated by transpiration in the plant and by perspiration in the animal. Submerged animals and plants are protected by the water, the temperature change of which depends upon the heat capacity of the particular body of water.

Treatment of biological materials with infrared radiations (780–1150 nm) either before or after x-ray treatment increases the rearrangements in chromosomes induced by the x-ray in tissues of plants and animals tested. The way in which the infrared radiations do this is unknown, but a comparable amount of conducted heat does not have the same effect.

Medical practitioners make use of infrared radiations to treat sprains, strains, bursitis, peripheral vascular diseases, arthritis, muscle pain, and many other pains for which heating gives relief, probably because of vasodilation of peripheral vessels. For this purpose, glow coil radiators are generally employed, but for some purposes, the more penetrating radiations (780–1400 nm) obtainable from incandescent sources are preferable to the glow coil to stimulate circulation deeper in the tissues. *See* BIOPHYSICS; THERMOTHERAPY.

Arthur C. Giese

Bibliography. J. R. Cameron and J. G. Skofronick, *Medical Physics*, 1978; W. R. Hendee, *Medical Radiation Physics*, 2d ed., 1979.

Infrared spectroscopy

The spectroscopic study of the interaction of matter with infrared radiation. Electromagnetic waves from the long-wavelength limit of visible light at 800 nanometers to the shortest microwaves at 1 mm are used. In the wave-number units usually employed (oscillations per centimeter, read as reciprocal centimeters), this corresponds to 12,500–10 cm^{-1}. *See* INFRARED RADIATION.

Infrared regions. The infrared is conveniently divided into three regions, according to the different types of atomic and molecular energy levels that can be studied and the experimental techniques.

Radiation in the near-infrared region, at 4000–12,500 cm^{-1} (wavelengths of 2.5–0.8 micrometers), can excite harmonics of molecular vibrations as well as low-energy electronic transitions in molecules and crystals. The usual source is a quartz-envelope tungsten-halogen lamp, with detection by photomultipliers or photodiodes (for the shorter wavelengths) and lead sulfide (PbS) or indium antimonide (InSb) photoconductors. *See* INCANDESCENT LAMP; PHOTOELECTRIC DEVICES.

The mid-infrared region, at 200–4000 cm^{-1} (50–2.5 μm), is where most stretching and bending fundamental vibrations of molecules occur. Blackbody sources (Nernst glowers and globars) are used. Detectors are thermal (thermocouples, bolometers), pyroelectric (deuterated triglycine sulfate, DTGS), or photoconductive InSb, doped germanium, and mercury cadmium telluride (HgCdTe). *See* BOLOMETER; HEAT RADIATION; PHOTOCONDUCTIVITY; PYROELECTRICITY; THERMOCOUPLE.

The far-infrared region, at 10–200 cm^{-1} (1000–50 μm), is where rotational transitions in gaseous molecules; certain low-frequency bending, torsional, and skeletal vibrations; and lattice modes in solids are observed. Plasma emission from a high-pressure mercury arc is the usual source; DTGS and doped-germanium bolometers are useful detectors, as is the Golay pneumatic cell. For still lower frequencies, microwave techniques are used. *See* CRYSTAL ABSORPTION SPECTRA; MICROWAVE SPECTROSCOPY; MOLECULAR STRUCTURE AND SPECTRA; RADIOMETRY.

Development. Although infrared radiation was discovered in 1800, the development of spectroscopy was slow until the 1880s. By then the bolometer detector had been invented and solar spectra recorded as far as 10 μm. The recognition that molecular vibrational frequencies are characteristic of specific chemical groups was emerging, based on studies in the near-infrared. Investigations in the early 1900s of the infrared spectra of organic compounds demonstrated their usefulness for identification, and established correlations between molecular structure and mid-infrared frequencies. The invention of the echelette or blazed grating, which concentrates most of the diffracted radiation into a single order, was a major experimental advance.

After the rapid development of infrared and electronics technology during World War II, recording infrared spectrometers became available commercially, and were widely adopted by industrial analytical laboratories in the 1940s. Initially, these used prisms, but were soon followed by higher-performance grating instruments of increasing sophistication. The development and commercialization of Fourier-transform spectrometers and tunable infrared lasers have further expanded the capabilities of infrared spectroscopy.

Instrumentation and techniques. The broad wavelength range of infrared radiation, and the few transparent optical materials available, require that infrared instruments be designed with reflecting optics: radiation is focused with front-surface aluminized mirrors rather than lenses. Because of strong infrared absorption by water vapor and carbon dioxide, operation takes place in a vacuum or the optical path is purged with dry nitrogen. Absorption spectroscopy is the principal method, where attention centers on the frequencies absorbed by the sample from a broadly emitting source. However, spectrometers and interferometers can easily be adapted to emission spectroscopy.

Dispersive spectrometers. In **Fig. 1**, infrared radiation from the source Q is focused by a spherical mirror M_2 onto the entrance slit S_1 of the monochromator, after passing through the sample cell SC. The beam is collimated by the off-axis paraboloid mirror M_3, dispersed by refraction through the prism P, and focused in the plane of the exit slit S_2 by a second reflection from M_3. A narrow spectral region of the dispersed radiation passes the exit slit and is focused by the ellipsoidal mirror M_7 onto the detector D, which converts the radiant energy into an electrical signal. Since the beam has been chopped at constant frequency by a rotating mechanical chopper C, this signal is an alternating current that is amplified by the lock-in amplifier A, controlling the pen of the chart recorder R. To scan the spectrum, M_4 is rotated by the drive mechanism DM, which also drives the recorder. Successive frequencies are thus moved across the exit slit, producing a record of signal intensity as a function of mirror position; with a proper mechanical linkage, this record is made linear in wavelength or wave number. The same arrangement can be used with a diffraction grating as the dispersive element; the prism is removed and mirror M_4 is replaced by the grating.

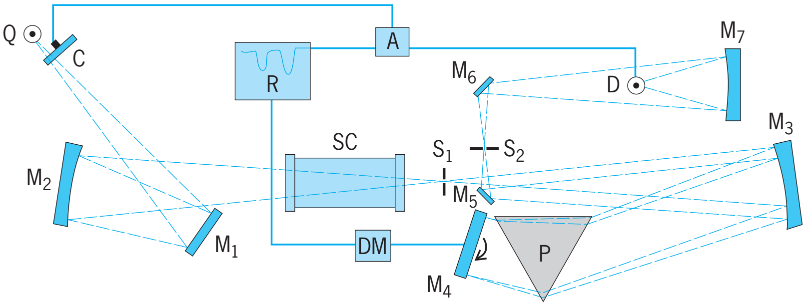


Fig. 1. Basic single-beam recording infrared prism spectrometer. The monochromator (the portion from the entrance slit S_1 to the exit slit S_2) is a Littrow mounting, a common arrangement for infrared instruments.

Even with an empty sample cell, the signal from such a single-beam instrument is affected by atmospheric absorption, variations in source output, absorption and scattering from the cell, and so forth. This is avoided with double-beam spectrometers, which use a beam-switching system of rotating mirrors to compare (typically some 10 times per second) the energy transmitted by the sample to that passing through a reference beam (which may contain an empty cell), and plot the ratio.

In the near-infrared, quartz prisms are used, and in the mid-infrared, alkali-metal halide or calcium fluoride (CaF_2) prisms, but no prism material is suitable beyond about 50 μm. Diffraction gratings can be used in all regions with the advantage, for equivalent optical configurations, of significantly higher resolving power than prisms. Prism instruments have resolutions of little better than 1 cm^{-1} near wavelengths of maximum dispersion, and much poorer than this elsewhere. Grating resolution can be several tenths of a reciprocal centimeter for commercial spectrometers; some specially built instruments can resolve a few hundredths of a reciprocal centimeter. In most laboratories these instruments are being replaced by other techniques, although inexpensive double-beam grating spectrometers are still manufactured. *See* DIFFRACTION GRATING; OPTICAL PRISM; RESOLVING POWER (OPTICS).

Fourier-transform spectroscopy. In a Michelson interferometer, light from the source strikes a thin, partially reflecting beam splitter at an angle of 45° and is divided into two beams that are returned by mirrors and recombined at the beam splitter. The intensity of the recombined and interfering beams, recorded as a function of the optical path difference (or retardation) as one mirror is moved, yields an interferogram. From this, the desired spectrum (that is, the source intensity as a function of wave number) can be recovered by the mathematical procedure known as a Fourier transform. *See* FOURIER SERIES AND TRANSFORMS; INTEGRAL TRANSFORM; INTERFEROMETRY.

The components of a Fourier-transform spectrometer (FTS) are shown in **Fig. 2**. Radiation from the source Q is chopped at C, collimated by mirror M_1, and divided at the beam splitter BS, with half sent to a fixed mirror M_2 and half to a movable mirror M_3. The recombined beams, with an optical-path difference $2L$, pass through a sample cell SC and are focused by M_4 onto the detector D. The signal is amplified (A) and stored in the digital computer DC. The interferogram is scanned by moving M_3 from zero path difference ($L = 0$) to some distance L_{max}; it is then Fourier-transformed and displayed as a spectrum on the recorder R. The spectral band pass of a Fourier-transform spectrometer is approximately $1/(2L_{max})$, and achieving a resolution of 0.1 cm^{-1}, for example, requires a mechanically very reasonable 5-cm mirror movement. At sufficiently high mirror velocities (about 1 mm/s for mid-infrared spectra), the signal is automatically modulated at audio frequencies and can be amplified directly, eliminating the chopper with its 50% signal loss. *See* DIGITAL COMPUTER.

Fourier-transform spectroscopy offers several advantages over dispersive methods; these are especially important for effective use of the limited radiant power available from most infrared sources. Whereas a spectrometer samples only one small frequency range at any given instant and must scan these frequencies sequentially to produce a spectrum, an interferometer processes information from all frequencies simultaneously (multiplex or Fellgett advantage). Furthermore, an interferometer passes a much greater light flux for a given resolving power than a spectrometer, which can accept only a very limited solid angle of source radiation because of the narrow slits required (throughput or Jacquinot advantage). These advantages can be translated into improvements of orders of magnitude in any one of the three interrelated important parameters of resolution, signal-to-noise ratio, and scan time. Another advantage is that the mirror movement is monitored to high precision with a fixed-frequency laser, so that the wave-number scale of the transformed spectrum is highly accurate compared with spectrometer readings.

Until the discovery of the Cooley-Tukey fast-Fourier-transform (FFT) algorithm in 1964, interferometric spectroscopy was largely impractical because of computational difficulties. The ready availability of inexpensive computers has now made it the accepted technique for high-performance infrared spectroscopy.

Commercial Fourier-transform spectrometers are marketed with resolutions from a few reciprocal centimeters for routine qualitative analyses to research instruments that can resolve better than 0.002 cm^{-1}. Typically, they operate from 4000 to 400 cm^{-1} on one beam splitter such as germanium-coated potassium bromide (KBr); this range can be extended broadly in either direction with different beam-splitter materials. For the far-infrared, Mylar films in various thicknesses are used. These instruments are controlled

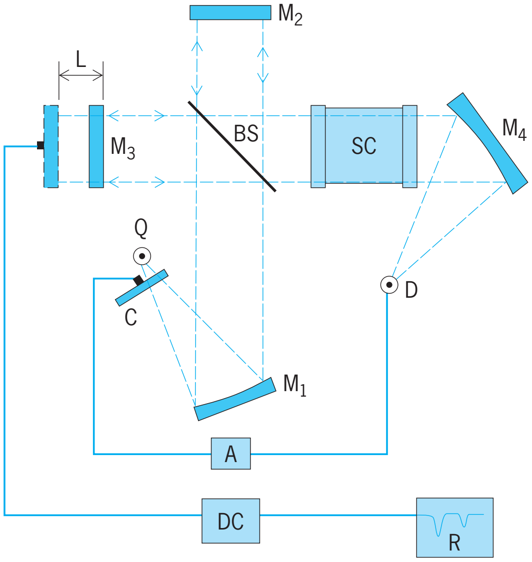


Fig. 2. Michelson interferometer used as a Fourier-transform spectrometer.

by microprocessors and include a digital computer to handle the Fourier transform. This computing power allows data manipulation such as repetitive scanning and signal averaging; background subtraction; spectral smoothing, fitting, and scaling; and searching digitized spectral libraries and databases to identify unknowns. Although backgrounds can be subtracted with software, some Fourier-transform instruments are designed for true optical double-beam operation. Many offer rapid-scan capability for the study of short-lived species and can be adapted to recording Fourier-transform Raman spectra. *See* MICROPROCESSOR; RAMAN EFFECT.

Tunable infrared lasers. In most experiments the lower limit to the linewidths of molecular transitions in gases is the Doppler broadening, the width caused by the translational (thermal) motion of the molecules. In the mid-infrared this is typically 10^{-2}–10^{-4} cm^{-1}; better resolution than even interferometers provide is desirable for spectroscopy of gases. Continuously tunable and nearly monochromatic laser sources, introduced about 1970, have met this need. *See* LASER; LASER SPECTROSCOPY.

Characteristics of some tunable lasers are listed in **Table 1.** Many of these emit beyond the ranges listed, which are limited to regions where useful high-resolution spectroscopy has actually been demonstrated. Other techniques, such as four-wave mixing, vibronic transition lasers, and polariton lasers, have been little used for spectroscopy.

Ideally, tunable infrared lasers require only a simple experimental arrangement: the source, a sample cell, and a detector with recorder. A low-resolution monochromator is often included to eliminate unwanted laser modes. In practice, many of these lasers continuously tune over very limited wavelength ranges, require constant adjustment, and tend to be complex and troublesome devices. However, they are indispensable when the highest resolution is required, and they have opened exciting new areas of infrared research.

If a tunable laser has enough power to saturate molecular transitions, as do carbon dioxide (CO_2) lasers, a technique known as saturation spectroscopy can be used to record spectra below the Doppler width, and thus exploit the full potential of these sources. This ultrahigh resolution can reveal a wealth of spectral detail that is otherwise lost in the Doppler broadening. *See* DOPPLER EFFECT.

Wave-number calibration. The accuracy of low-resolution instruments can be quickly checked by recording the spectra of standards such as polystyrene film or a solution of camphor and cyclohexanone in indene; these provide sharp absorption peaks with positions known within ±0.5 cm^{-1}. For high-resolution spectra, tabulations are available giving the frequencies of thousands of rotation-vibration lines of light gaseous molecules such as hydrogen chloride (HCl), carbon monoxide (CO), water vapor (H_2O), hydrogen cyanide (HCN), carbon dioxide (CO_2), nitrous oxide (N_2O), ammonia (NH_3), and ethylene (C_2H_2), with accuracies of ±0.005 cm^{-1} or better.

Sample handling. Gases are examined in simple cylindrical cells with infrared-transmitting windows; a typical path length is 10 cm. For trace species or weak absorptions, compact multiple-reflection White cells (named after their inventor) are made with adjustable optical path lengths up to 100 m (300 ft) and more. Liquid cells are available for sample thicknesses of 0.01 mm to several centimeters. Solids can be observed directly in thin sections; as mulls in mineral or fluorocarbon oils; as finely ground dispersions in disks of an infrared-transparent salt such as potassium bromide; or in solutions of solvents such as carbon tetrachloride (CCl_4) or carbon disulfide (CS_2) that have few infrared bands. (The usefulness of water is limited by its strong, broad infrared absorptions.)

Many accessories have been designed for sampling problems. Infrared microscopes with reflecting optics can focus on objects as small as 10 μm to record the spectra of subnanogram samples; this is useful in particulate and fiber analysis and forensic work. Attenuated total reflection, where the infrared beam is transmitted inside a thin, highly refractive crystal [such as zinc selenide (ZnSe) or germanium] by multiple internal reflections, and is

TABLE 1. Tunable mid-infrared laser sources*

Device	Demonstrated spectral coverage, μm	Maximum resolution, cm^{-1}	Typical continuous-wave power, W
Semiconductor diode lasers	3–27	2×10^{-6}	10^{-3}
Gas lasers			
Waveguide CO_2	9–11†	3×10^{-7}	1
Zeeman-tuned neon or xenon	~3.5†	1×10^{-3}	10^{-3}
	7–8†	1×10^{-3}	?
Spin-flip Raman lasers	5–6	1×10^{-6}	0.1
Color-center lasers	1.5–4	3×10^{-8}	10^{-2}
Optical parametric oscillators	1–16	$<1 \times 10^{-3}$	10^{-3}
Nonlinear optical mixing techniques			
Difference-frequency generation	2–6	1×10^{-4}	10^{-6}
Tunable-sideband generation	~3.4†	2×10^{-10}	?
	9–11†	3×10^{-6}	10^{-3}

*Adapted from R. S. McDowell, Infrared laser spectroscopy, *Proc. SPIE*, 380:494–507, 1983.
†Tunable only near discrete lines in these regions.

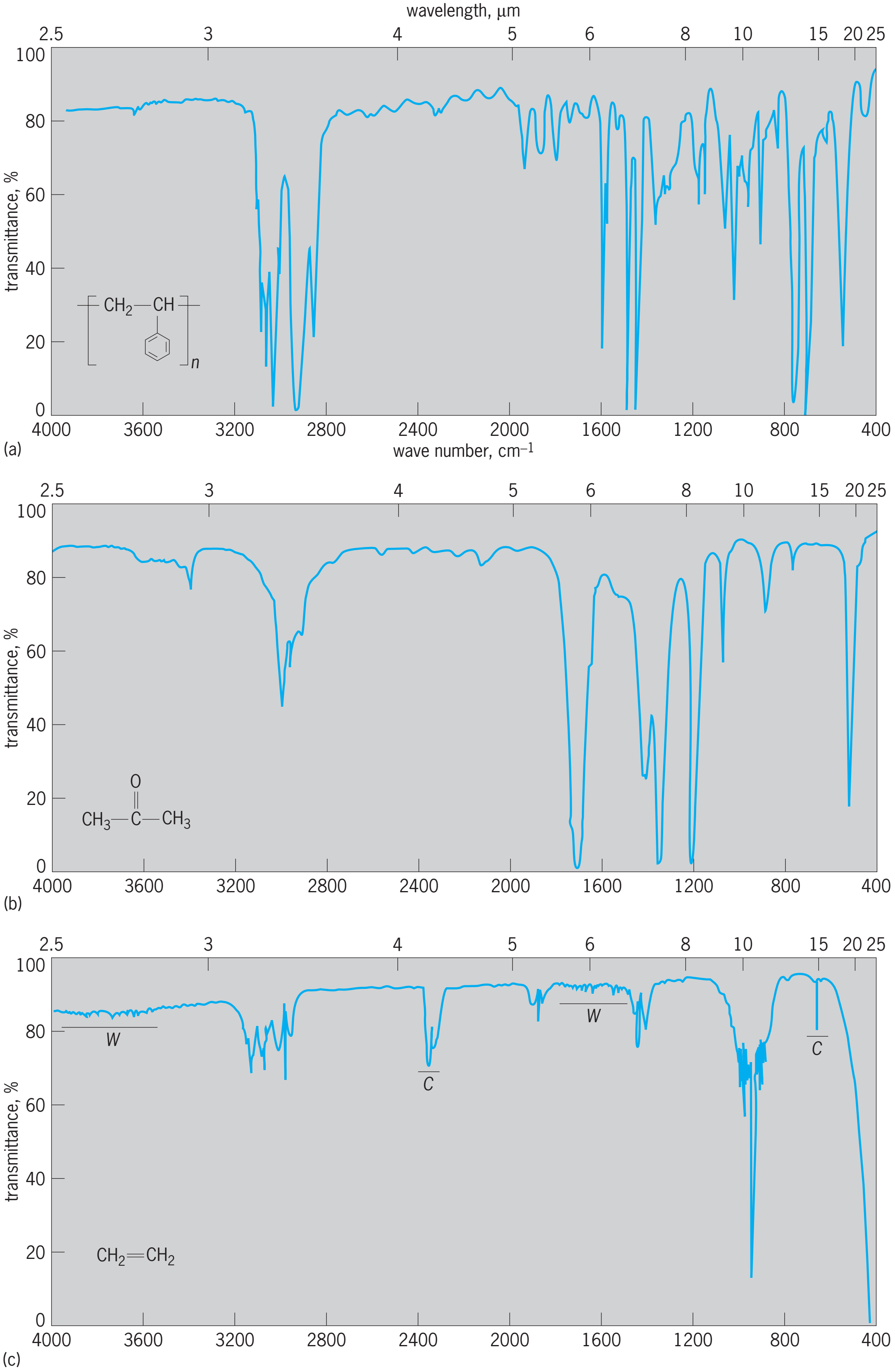


Fig. 3. Mid-infrared absorption spectra recorded on a Fourier-transform spectrometer at a resolution of 1 cm^{-1}. (*a*) Polystyrene film, 0.05 mm thick. The weak intensity fluctuations between 2000 and 2800 cm^{-1} are interference fringes, caused by reflections from the surfaces of the plastic, and can be used to determine the film thickness. (*b*) Liquid acetone, capillary film between cesium iodide (CsI) plates. (*c*) Ethylene gas at about 30 torr (4 kilopascals) in a 10-cm (4-in.) cell. The absorptions marked W and C are due to traces of water vapor and carbon dioxide (CO_2), respectively. Below 600 cm^{-1} the potassium chloride (KCl) windows absorb. (*V. T. Hamilton, Los Alamos National Laboratory*)

absorbed by a sample in optical contact with the crystal, may be appropriate for difficult materials such as gels and strongly absorbing liquids. Photoacoustic detectors can be used for nearly opaque samples. Other accessories provide spectra of specularly reflected light from coatings and films; of diffuse reflectance from inhomogeneous samples such as powders; of samples under high pressure in diamond-anvil cells; and so forth. Many devices are available for holding samples at cryogenic or elevated temperatures. *See* HIGH-PRESSURE PHYSICS; PHOTOACOUSTIC SPECTROSCOPY.

Typical spectra. Infrared spectra are usually plotted as percent transmittance T or absorbance A on a scale linear in wave number ν (less commonly, in wavelength λ). Transmittance is the ratio of the intensity of radiation transmitted by the sample (I) to that incident on the sample (I_0), expressed as a percentage, so that $T = 100I/I_0$; this is related to absorbance by Eq. (1). Transmittance and absorbance,

$$A = \log_{10}\left(\frac{I_0}{I}\right) = \log_{10}\left(\frac{100}{T}\right) \qquad (1)$$

of course, vary with frequency and should formally be expressed as $T(\nu)$ and $A(\nu)$.

Examples of mid-infrared spectra recorded on a commercial instrument are shown in **Fig. 3.** The very strong absorption in Fig. 3*b* (liquid acetone) at 1705–1725 cm^{-1} is characteristic of the carbonyl $C{=}O$ stretch in saturated aliphatic ketones. The rotational lines in Fig. 3*c* (ethylene gas) between 850 and 1050 cm^{-1} arise from quantized changes in the amount of rotational excitation that occurs simultaneously with the $=CH_2$ out-of-plane "wagging" vibrational transition at 949 cm^{-1}. A similar structure is seen on the C—H stretch between 3050 and 3200 cm^{-1}, and would be evident in all bands at higher resolution. In liquids and solids, this rotational structure is suppressed.

A single rotation-vibration band of a gaseous molecule, the infrared-active stretch (designated ν_3) of sulfur hexafluoride (SF_6), is shown in **Fig. 4** at four resolutions from 2 to 10^{-6} cm^{-1}, to illustrate the additional detail revealed at higher resolution. The emission linewidth of the tunable semiconductor diode laser used to record Fig. 4*c* was less than 10^{-5} cm^{-1}, but the effective resolution is the sulfur hexafluoride Doppler width of 0.0010 cm^{-1}. Figure 4*d* is a sub-Doppler saturation spectrum taken inside the linewidth of a carbon dioxide (CO_2) laser line at an effective resolution of less than 10^{-6} cm^{-1}. This was recorded in the derivative mode, where the ordinate is $dA/d\nu$, the derivative of A with respect to ν, instead of A or T. Many thousands of features like those in Fig. 4*c* and *d* have been assigned to specific sulfur hexafluoride transitions.

Applications. Among the more important applications of infrared spectroscopy are chemical analysis and the determination of molecular structures.

Qualitative analysis. Infrared spectra are ideal for identifying chemical compounds because every molecule [except homonuclear diatomics such as nitrogen (N_2), oxygen (O_2), and chlorine (Cl_2)] has an infrared spectrum. Since the vibrational frequencies depend upon the masses of the constituent atoms and the strengths and geometry of the chemical bonds, the spectrum of every molecule is unique (except for optical isomers). Pure unknowns can thus be identified by comparing their spectra with recorded spectra; catalogs are available in digitized versions, and searches can be made rapidly by computer.

Simple mixtures can be identified with the help of computer software that subtracts the spectrum of a pure compound from that of the unknown mixture. More complex mixtures may require fractionation

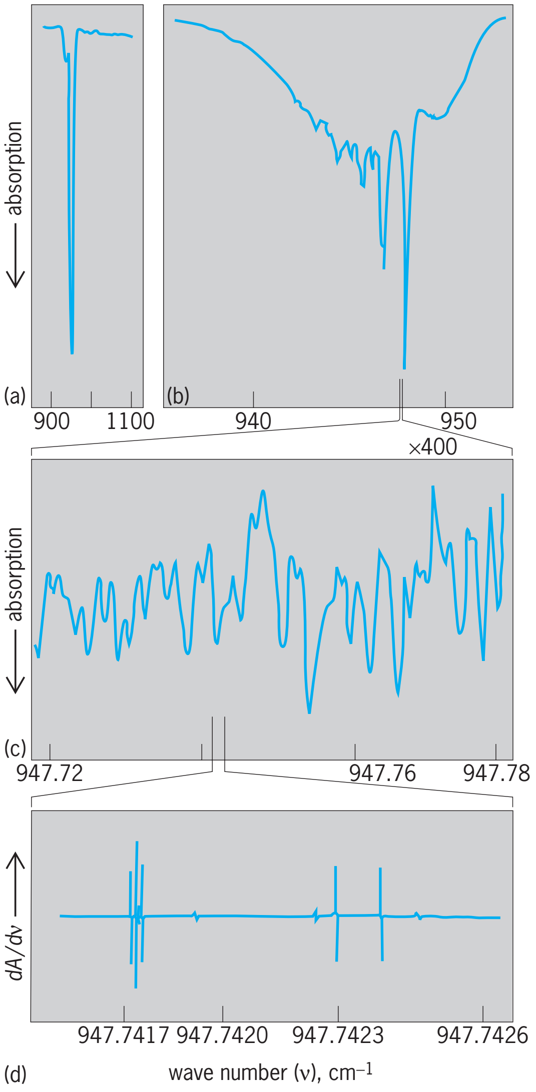


Fig. 4. S—F stretching fundamental ν_3 of sulfur hexafluoride (SF_6) as it appears with increasing resolving power. (*a*) Fourier-transform spectrum at resolution of 2 cm^{-1}. (*b*) Fourier-transform spectrum at resolution of 0.06 cm^{-1}. (*c*) Expansion of small portion of the strong central peak in part *b*, recorded with a tunable semiconductor diode laser at an effective resolution of 0.0010 cm^{-1}. (*d*) Further resolution of the absorption line at 947.742 cm^{-1} into five strong components with sub-Doppler saturation spectroscopy at an effective resolution of $<10^{-6}$ cm^{-2}. (*After R. S. McDowell, in J. R. Durig, ed., Vibrational Spectra and Structure, vol. 10, pp. 1–151, 1981*)

TABLE 2. Group frequencies for O—H, N—H, and C—H stretches*

Functional group	Frequency, cm^{-1}
Alcohol and phenol —OH	3593–3643 (medium)
Carboxylic acid —CO—OH	3500–3560 (medium)
Most NH	3310–3540 (medium)
Hydrazine >N—NH_2	3139–3327, 3354–3427 (medium)
Amino acid —NH_3^+	3030–3130 (medium)
Alkane —CH_3 (methyl)	2872 ± 10, 2962 ± 10 (strong)
Alkane >CH_2 (methylene)	2853 ± 10, 2926 ± 10 (strong)
Alkane —CH< (methylidyne)	2890 ± 10 (weak)
Alkene —CH═CH_2 (vinyl)	3010–3040, 3075–3095 (medium)
Alkene >C═CH_2 (vinylidene)	3075–3095 (medium)
Alkene —CH ═ (methine)	3010–3040 (medium)
Alkyne —C≡CH (ethynyl)	3315 5 (medium)
Aldehyde —CH═O (formyl)	2700–2900 (weak; usually 2)
Aromatic CH	~3030 (medium), ~3070 (weak)
Heterocyclic (azine) CH	3010–3050, 3060–3090 (weak)

*Hydrogen bonding lowers the OH and NH frequencies by 100–1000 cm^{-1} from these values, which apply to free molecules.

SOURCE: After L. J. Bellamy, *The Infra-Red Spectra of Complex Molecules,* vol, 1, 3d ed., 1975.

first. This has led to the development of combinations of analytical techniques, such as gas chromatography used together with Fourier-transform infrared spectroscopy; instruments that combine these functions are available commercially.

Many functional groups have characteristic infrared frequencies (**Table 2**) that are relatively independent of the molecular environment. Often, specific conclusions can be drawn from frequencies, and it may be possible to identify even a new compound from its spectrum. Group frequencies are most useful above about 1500 cm^{-1}; below this the absorptions are due to, or are influenced more by, the skeletal vibrations of the molecule. This is the "fingerprint" region, where even similar molecules may have quite different spectra.

With the help of Table 2 and the spectrum of Fig. 3*a*, the absorptions at 2851 and 2924 cm^{-1} can be immediately assigned to methylene. That at 3027 cm^{-1} could be alkene, heterocyclic, or aromatic C—H; a more extensive table of group frequencies shows that the last is correct, from the phenyl C—C ring "breathing" stretches at 1400–1650 cm^{-1} and out-of-plane C—H bends at 600–800 cm^{-1}. The latter bands are characteristic of monosubstitution, as is the pattern of weak combinations between 1650 and 2000 cm^{-1}. There is no indication of other functional groups. This evidence suggests polystyrene, and the identification would be confirmed by comparison with a known spectrum. *See* QUALITATIVE CHEMICAL ANALYSIS.

Quantitative analysis. Quantitative infrared analysis is possible because the absorbance defined in Eq. (1) is directly proportional to the number of absorbing molecules, and in mixtures is the sum of the absorbances of the individual components at the frequency of interest. The Bouguer-Lambert-Beer law is given in Eq. (2), where $a(\nu)$ b, and c are the absorptivity (a frequency-dependent molecular property), path length, and concentration, respectively. When the absorptivity of a compound is known at a given frequency, for example, at an absorption peak, the concentration of that compound in a mixture can be determined from Eq. (2). Mixtures with n components can be analyzed by measuring absorbances at n selected frequencies and solving the resulting system of simultaneous linear equations.

$$A(\nu) = a(\nu)\, bc \qquad (2)$$

Traditional infrared analysis, although highly specific, is not notably sensitive: typical detection limits are 0.1–1% with precisions of a few percent of the quantity measured, except in especially favorable cases. These limits improve with increasing resolution, until the resolution becomes somewhat less than the width of the feature being measured. Tunable-laser sources offer much greater sensitivity because of their narrow linewidths and easy adaptability to first- and second-derivative detection (see also Fig. 4*d*). Lasers are used for concentration measurements of trace atmospheric gases and pollutants; detection limits in the parts-per-trillion range have been reported. *See* QUANTITATIVE CHEMICAL ANALYSIS.

Molecular structure. Many details of molecular structure and dynamics can be extracted from an infrared spectrum, especially for light molecules that can be examined in the gas phase and therefore exhibit rotational structure. The near-infrared absorption of acetylene (H—C≡C—H) in **Fig. 5** provides an example. This is a combination band, in which two fundamentals are excited, ν_1 (symmetric C—H stretch) at 3373 cm^{-1} and ν_5 (bend) at 730 cm^{-1}. The difference between the band center at 4091 cm^{-1} (indicated by A in Fig. 5) and the sum of $3373 + 730 = 4103$ cm^{-1} is a measure of the anharmonicity of the vibrations involved.

The 3:1 intensity alternation in the rotational lines (B) is due to nuclear spin statistics, and indicates that acetylene is linear and symmetric. This molecular symmetry can also be deduced from the number of fundamental vibrations in the infrared and Raman spectra, and the number of infrared–Raman coincidences. From the spacing of the rotational lines (C), the molecular moment of inertia can be obtained; this and the moment of the deuterated species yield the bond lengths, 0.106 nm for C—H and 0.121 nm for C≡C. The small change in the rotational line spacing across the band is due to vibration–rotation

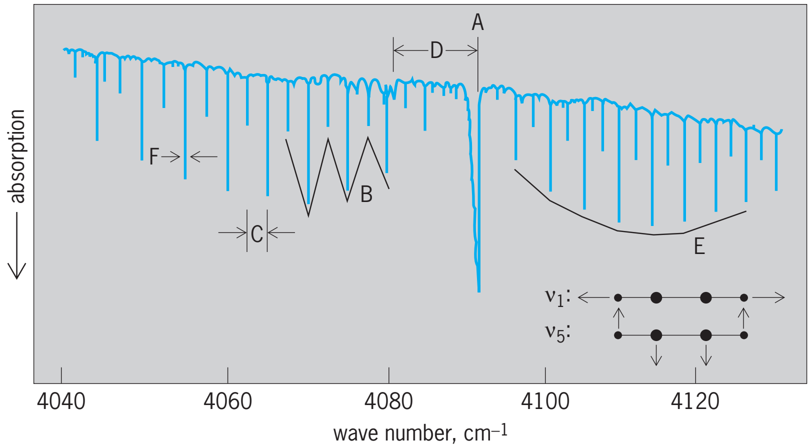


Fig. 5. Spectrum of the $\nu_1 + \nu_5$ rotation–vibration band of acetylene recorded with a grating spectrometer, resolution 0.05 cm^{-1}; gas at 25°C (77°F) and 100-torr (13.3-kPa) pressure, 25-cm (10-in.) path length. The letters indicate features that are discussed in the text. (*After A. R. H. Cole, Tables of Wavenumbers for the Calibration of Infrared Spectrometers, 2d ed., Pergamon Press, 1977*)

interaction, that is, to the fact that the bond lengths differ slightly in the ground and excited vibrational states.

The weak absorption at 4080 cm^{-1} is the center of the corresponding band of the isotopic species $H^{12}C^{13}CH$, present in its natural abundance of 2.2%. From this frequency shift (D), and the shifts for the deuterated species, the molecular force constants can be calculated, which are measures of the resistance of the chemical bonds to deformations. For acetylene, the stretching force constants are 0.64 micronewton/nm (6.4 millidyne/Å) for the C—H bond and 1.63 μN/nm (16.3 mdyn/Å) for C≡C, which are typical values for single and triple bonds, respectively. Knowledge of the bond strengths helps in understanding the electronic structure.

Finally, the spectrum provides information on the conditions of the sample. The overall intensity contour of the rotational lines (E) is characteristic of the temperature of the gas; the individual linewidths (F) increase with pressure over their minimum (Doppler) value. Therefore, good estimates of temperature and pressure can be made if these are not otherwise known. The usefulness of such information in remote sensing is obvious.

Other applications. Other applications include calculation of thermodynamic quantities from vibrational and rotational energy levels; studies of intermolecular forces in condensed phases; distinguishing between free or hindered internal rotation (as of methyl groups); quantitative intensity measurements to obtain bond dipole moments; studies of molecular interactions in adsorption, surface chemistry, and catalytic processes; time-resolved monitoring of transient species and chemical reaction kinetics; characterization of reactive molecules isolated at cryogenic temperatures in rare-gas lattices (matrix isolation); electronic energy states in semiconductors and superconductors; studies of biological molecules and membranes (effects of hydrogen bonding); and analysis of energy levels in laser systems. Double-resonance techniques, in which molecules are excited by an intense pulse of ultraviolet, infrared, or microwave energy, and simultaneously probed by infrared or microwave spectroscopy, are useful in monitoring population changes in molecular energy levels, energy transfer and relaxation, and multiphoton absorption processes. *See* ADSORPTION; CATALYSIS; CHEMICAL DYNAMICS; HYDROGEN BOND; INTERMOLECULAR FORCES; SEMICONDUCTOR; SPECTROSCOPY; SUPERCONDUCTIVITY. **Robin S. McDowell**

Bibliography. J. Chalmers and P. Griffiths (eds.), *Handbook of Vibrational Spectroscopy*, 5 vols., 2002; A. A. Chirstty, Y. Ozaki, and V. G. Gregoriou, *Modern Fourier Transform Infrared Spectroscopy*, 2001; N. B. Colthup, L. H. Daly, and S. E. Wiberley, *Introduction to Infrared and Raman Spectroscopy*, 3d ed., 1990; H.-U. Gremlich and B. Yan, *Infrared and Raman Spectroscopy of Biological Materials*, 2000; B. Schrader, *Infrared and Raman Spectroscopy: Methods and Applications*, 1994; J. I. Steinfeld, *Molecules and Radiation*: *An Introduction to Modern Molecular Spectroscopy*, 2d ed., 2005; B. H. Stuart, *Infrared Spectroscopy: Fundamentals and Applications*, 2004.

Infrasound

Low-frequency sound waves that cannot be sensed by the human ear. Infrasound was first detected on barometers around the world following the eruption of the volcano Krakatoa in 1883. We now know subaudible sounds are generated by a wide variety of natural phenomena, such as large storms and meteors, and some human-produced sources, such as large chemical or nuclear explosions. Infrasound is particularly useful for basic research, for example, in the study of atmosphere phenomena; as it decays slowly across thousands of kilometers. How infrasound is generated, how it propagates through our turbulent atmosphere, and how it can be recorded amid noise due to air circulation are the subjects of ongoing research.

Origin. Humans can hear sounds from 20 to 20,000 Hz, in the middle of the acoustic spectrum. Infrasonic signals (below 20 Hz) are subaudible. Infrasound is generated by myriad sources, either impulsive, that is, relatively short-lived, or ongoing. To generate infrasound, these sources must be energetically significant and move a large volume of air. Some sources, such as hurricanes and earthquakes, are very large. Earthquakes compress the air and excite infrasound waves by vibrating large areas of the Earth's surface in an up-down pistonlike fashion or by shaking nearby topography. Volcanoes, tornadoes, avalanches, surf, aurora, sprites, projectiles (such as supersonic aircraft, rockets, and meteors), and explosive events (such as atmospheric nuclear tests and surface mine blasts) also generate infrasound waves. Some large animals, including elephants, rhinoceros, and whales, are known to use infrasound to communicate over distances of many kilometers.

The frequency of infrasound waves scales with the size of the source, with the largest sources generating the lowest-frequency signals (**Fig. 1**). Thunder and sonic booms, for example, generate

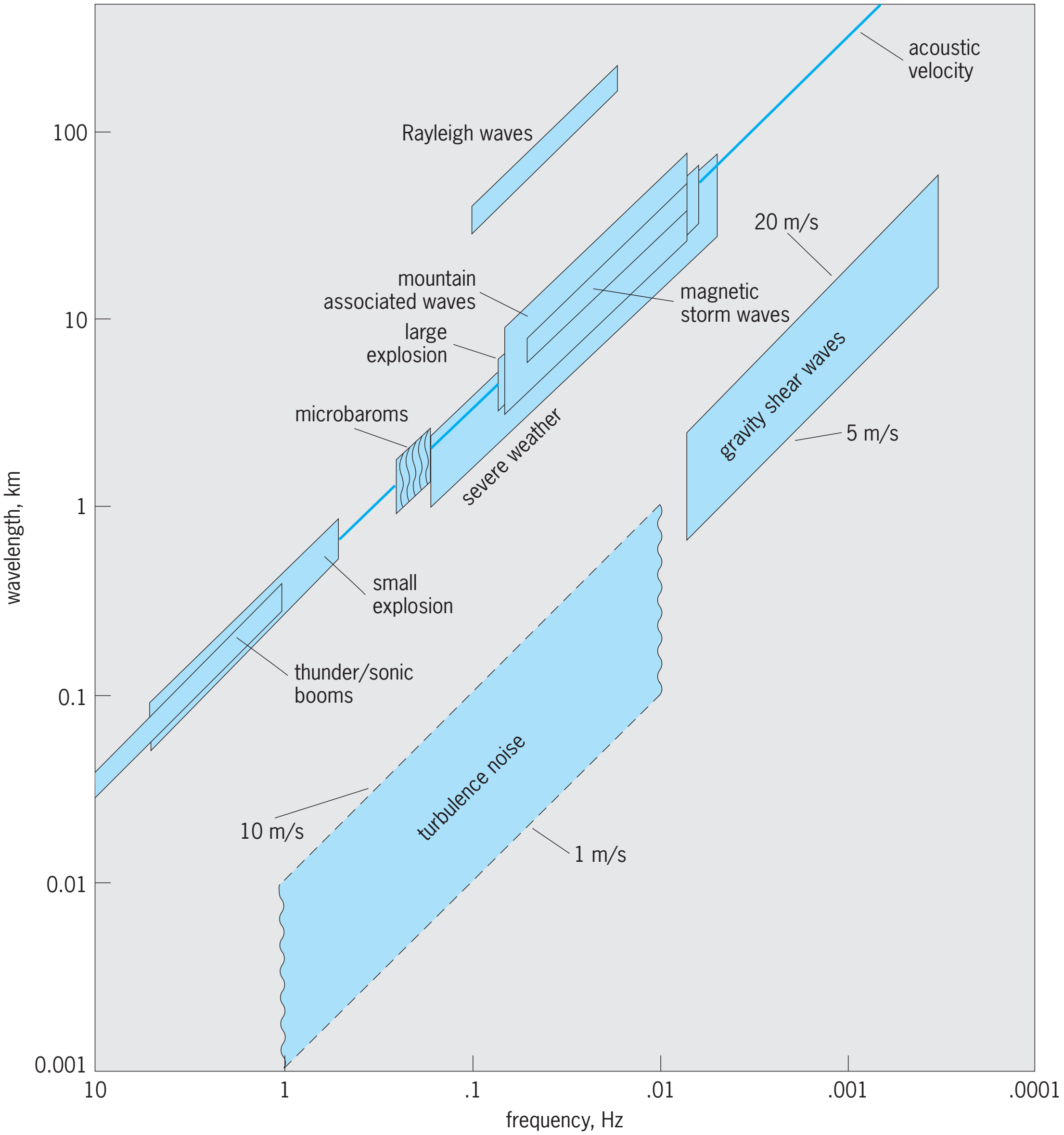


Fig. 1. Wavelength of subaudible acoustic waves as a function of frequency. Turbulent noise exists at the same frequency as acoustic signals, but at much shorter spatial scales. (*Adapted from Alfred Bedard with permission*)

higher-frequency infrasound waves. Lower-frequency infrasound is generated by larger sources, including large explosions and severe storms. Very low frequency infrasound is due to turbulence from air flowing over mountains. Mountain-associated waves are due to turbulence from air flowing over mountains. Earth's gravity becomes important at very low frequencies (below ∼0.01 Hz). *See* TURBULENT FLOW.

Propagation. Infrasound is a mechanical wave phenomenon similar to seismic waves that travel through the solid Earth. The propagation of infrasound waves, however, differs significantly from seismic waves in that the infrasound velocity structure of the atmosphere changes rapidly. As with any wave phenomenon, infrasound refracts, or changes direction, due to changes in the infrasound velocity of the medium through which it passes. Infrasound refracts away from regions of high velocity. The velocity of infrasound is directly proportional to the square root of the absolute temperature of air, and increases in the direction of wind due to advection. Both temperature and wind depend on altitude, geographic location, time of day, and time of year. *See* REFRACTION OF WAVES; SEISMOLOGY.

To be useful for research, infrasound must return to sensors on the ground. With favorable winds, infrasound is efficiently refracted back to the Earth's surface from within the stratosphere (**Fig. 2**). Infrasound returns to the Earth's surface from within the thermosphere due to the very significant increase in temperature with increasing altitude. Temperature decreases with increasing altitude in the troposphere, the lowest atmospheric layer; however, infrasound can be refracted back to the ground from within this layer in the presence of strong temperature inversions or if the infrasound propagates in the direction of strong winds. *See* ATMOSPHERE; ATMOSPHERIC ACOUSTICS.

Infrasound is valuable for monitoring nuclear

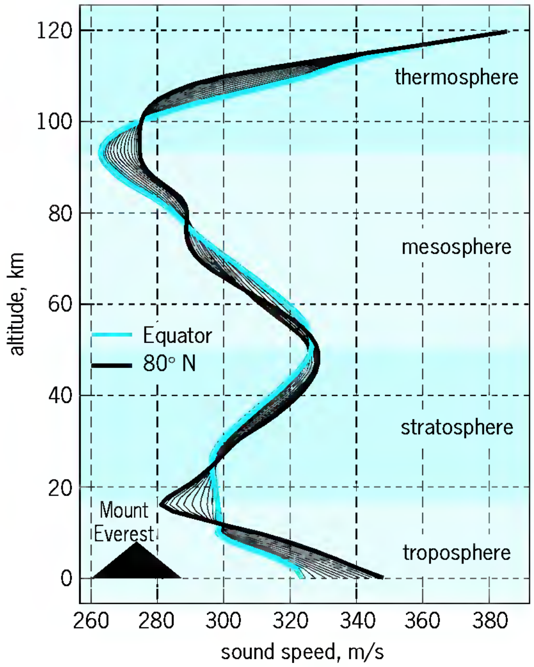


Fig. 2. Annual mean sound speed (Equator to 80° N). The speed of sound tracks air temperature and varies with altitude. Sound speed decreases with increasing altitude in the troposphere as the air temperature decreases away from Earth's surface. Sound speed increases in the stratosphere, decreases in the mesosphere, and increases sharply with increasing altitude in the thermosphere. Sound waves refract away from regions where the sound velocity is high, and thus turn back to the Earth's surface within the thermosphere and, with favorable winds, from within the stratosphere. Waves ducted within the troposphere have been observed and are believed to be due to temperature inversions or unusually strong winds in the direction of propagation of the sound waves.

testing and other atmospheric phenomena, as it propagates through the atmosphere with relatively little dissipation. The rate at which energy is lost scales with frequency, with the highest frequencies attenuated relatively rapidly. Infrasound waves above and below 1 Hz are termed near- and far-infrasound, respectively.

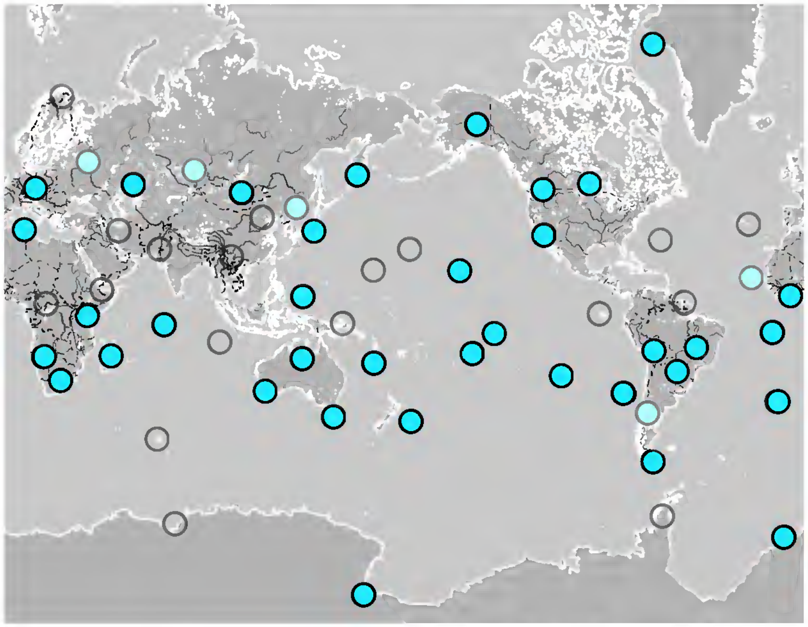


Fig. 3. Global infrasound network. Of the 60 planned operating stations, over half are built and recording data (dark color circles), several are under construction (light color circles), and a third have not yet been built (open circles).

The use of infrasound for the study of atmospheric phenomena requires that the dynamic structure of the atmosphere be taken into account. Atmospheric models based on climatologies, numerical weather prediction models, and pinpoint measurements, such as radiosondes, are used to synthesize the propagation of infrasound.

Detection. It is possible to feel infrasound waves near a particularly intense source, such as a jet airplane on take-off. The eruption of Krakatoa generated the loudest sound in recorded history. Infrasound waves from this cataclysmic event registered several times on barometers worldwide due to repeated circuits around the globe. This was highly unusual, as infrasound waves, in general, are very faint and can be detected only with sensitive instruments or indirectly through the effect on structures.

The ability to detect infrasound waves is limited by noise from atmospheric turbulence. Most infrasonic noise in the frequency band of interest to infrasound researchers is coherent across distances of less than several meters, while infrasonic signals can be coherent over distances in excess of 100 m (330 ft; Fig. 1). The ratio of coherent signal to incoherent noise can be increased by averaging the infrasonic pressure field over an area smaller than the coherence limit of signals but greater than the coherence limit of noise. The most common method used to incoherently sum atmospheric noise while coherently summing the signal is to sample atmospheric pressure at many points distributed spatially (across an area ranging up to 100 m), transmitting the sound through an array of pipes to a microbarometer where they are summed acoustically. *See* ACOUSTIC NOISE; COHERENCE.

Some of the drawbacks resulting from the mechanical reduction of noise can be avoided by integrating sound waves at the speed of light. One such system under development uses fiber-optic cables for this purpose. Another approach is to record infrasound using a spatial grid of instruments, and then sum the data from the instruments optimally to reject noise while preserving coherent signals. *See* OPTICAL FIBERS.

Global monitoring network. A global infrasound network under construction (as of mid-2006) will comprise 60 continuously operating arrays (**Fig. 3**). The network is designed to provide uniform coverage of the Earth's atmosphere, given the constraint that all arrays must be located on land. Each array in the network will comprise four to eight microbarometers with apertures of 1–3 km (0.6–1.8 mi). Each microbarometer in an array is equipped with a noise-reducing filter averaging noise due to turbulence at spatial scales up to about 70 m (230 ft). The use of arrays of sensors, rather than single instruments, allows the detection of more faint infrasound signals and estimation of the speed and direction of the incident waves. The network is designed for optimal detection of signals between 0.1 and 1.0 Hz.

Array processing. Array processing methods take advantage of the coherence, or similarity, of a signal recorded at different elements in an array to provide information about the signal, such as the velocity at which the signal crossed the array. The velocity can be used to infer the direction to the source and the path that the signal took through the atmosphere. The most commonly used method in the infrasound field is based on the progressive multichannel correlation (PMCC) algorithm. The PMCC method is based on the cross-correlation of data from different elements in an array. Cross-correlation is a mathematical measure of similarity of two segments of data. If two elements in an array record the same signal, the cross-correlation procedure will provide an estimate of how similar the recordings are, and how much the signal is shifted in time between the two recordings due to the propagation delay. The PMCC method extends this idea to subarrays of several elements and defines a closure relation as a more demanding test of whether a signal is present in the data. *See* SIGNAL PROCESSING; STATISTICS.

Consider a subarray of three elements (i, j, k). The PMCC method defines signal consistency (r_{ijk}) for the subarray as

$$r_{ijk} = \Delta t_{ij} + \Delta t_{jk} + \Delta t_{ki}$$

where Δt_{ij} is the time delay between the arrival of a signal at sensors i and j. A signal is considered consistent if the sum of the time delays is below a predefined threshold. The PMCC method applies this test to several small subarrays within an array. If a highly correlated and consistent signal is found, the test is repeated for progressively larger subarrays to refine the estimate of the velocity of the signal across the array. This procedure is repeated in several frequency bands and in numerous overlapping windows spanning the time of interest. The use of small subarrays avoids ambiguity problems inherent in correlating distant signals.

A single array recording can be a very accurate estimate of the direction to the source of the signal. The velocity of the signal across the array can be used to infer the path through the atmosphere taken by the infrasound; however, pinpointing the location of the source requires recordings made by several arrays.

Continuing basic research. In monitoring the atmosphere continuously, the infrasound network offers an unprecedented opportunity to better understand human-produced and natural atmospheric phenomena on a global scale. Research continues on how sound propagates through our unsteady atmosphere and how clear recordings of distant events can be made despite noise due to atmospheric turbulence. Researchers are improving models of the atmosphere and are collecting information about significant atmospheric sources to provide a basis for this research. *See* SOUND.

Michael A. H. Hedlin

Bibliography. A. J. Bedard and T. M. Georges, Atmospheric infrasound, *Phys. Today*, March, pp. 32–37, 2000; M. A. H. Hedlin et al., Listening to the secret sounds of Earth's atmosphere, *Eos Trans. Amer. Geophys. Union*, 83:557, 564–565.

Inhibitor (chemistry)

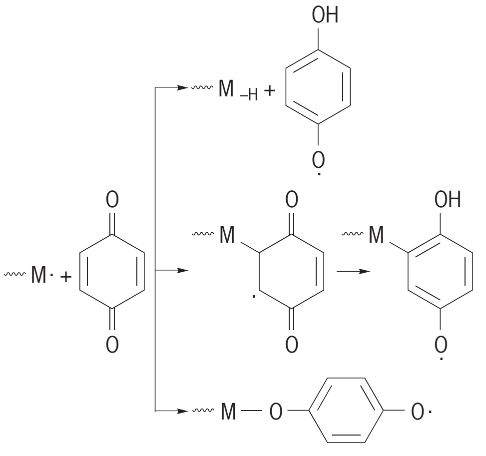
A substance which is capable of stopping or retarding a chemical reaction. To be technically useful, such compounds must be effective in low concentrations, usually under 1%. The type of reaction which is most easily inhibited is the free-radical chain reaction. The study of inhibitor action is often used as a diagnostic test for free-radical chain character of a reaction. Vinyl polymerization and autoxidation are two important examples of the class. Another reaction type for which inhibitors have been found is corrosion, particularly in aqueous systems. The economic importance of corrosion inhibition can scarcely be overestimated. An understanding of inhibitor action depends on an understanding of the processes to be interrupted.

Inhibition of vinyl polymerization. This type of inhibitor action must be considered in terms of the accepted mechanism for the polymerization process, which may be summarized as reaction sequence (1).

$$\begin{array}{rcll} \text{P} & \longrightarrow & 2\text{R}\cdot & \left.\begin{array}{l}\\ \\ \end{array}\right\} \text{initiation} \\ \text{R}\cdot + \text{M} & \longrightarrow & \sim\text{M}\cdot & \\ \sim\text{M}\cdot + \text{M} & \longrightarrow & \sim\text{M}\cdot & \text{propagation} \\ 2\sim\text{M}\cdot & \longrightarrow & \multicolumn{2}{l}{\text{termination by dimerization or}} \\ & & \multicolumn{2}{l}{\text{disproportionation}} \end{array} \qquad (1)$$

The symbol P represents a catalyst, often a peroxide, R· is a free radical derived from the catalyst, M is a monomer, and ~M· is a growing polymer chain. The polymerization will be stopped or inhibited if some added substance (an inhibitor) reacts more readily than does the monomer with R· to yield a product which will not sustain the polymerization. Every reaction chain is stopped until the inhibitor is consumed. If the added substance (a retarder) is somewhat less reactive, the monomer can compete more successfully for the initiating radicals, so that the result is retardation rather than total inhibition. The difference between inhibition and retardation is one of degree rather than of kind.

Phenolic compounds and quinones. These interact with an initiating radical or a growing polymer chain either by hydrogen atom abstraction or by radical addition to an unsaturated linkage. These interactions are represented by sequence (2).



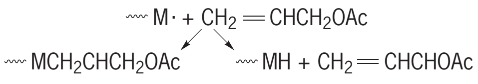
(2)

The phenoxy radicals produced are stabilized by resonance and are not sufficiently reactive to add to the vinyl linkage of another monomer molecule. The usual fate of these radicals is further hydrogen atom loss by reaction with a second polymer radical or by disproportionation with another phenoxy radical to yield a quinone and a hydroquinone, both of which may continue to act as inhibitors. All the reaction possibilities shown have been demonstrated experimentally. Although the efficiencies of phenols and quinones for the interruption of polymerizations vary with the structures of these molecules, they may be classed as inhibitors. Aromatic amines react in a similar manner.

Nitroaromatics. As typified by trinitrobenzene, nitroaromatics function as retarders rather than as inhibitors in most polymerizations. It is necessary, however, to consider the specific reaction involved. Thus polynitroaromatics inhibit the polymerization of vinyl acetate, retard that of styrene, and have no effect on that of methyl methacrylate. No clear-cut mechanism has been established for the interaction of nitro compounds with free radicals.

Monomers. Both monomers and the radicals derived from them differ greatly among themselves in reactivity. Thus, although certain monomers may copolymerize with one another, others may actually function as inhibitors. Styrene and vinyl acetate, for example, both polymerize well when alone. Styrene, however, inhibits the polymerization of vinyl acetate. This occurs because the vinyl acetate radical and the styrene monomer are highly reactive, whereas the styrene radical and the vinyl acetate monomer are not. A small amount of styrene added to vinyl acetate will rapidly react with any vinyl acetate radicals formed when polymerization is initiated. The resulting styrenelike radical will react only very slowly with the vinyl acetate monomer. In the overall process the chain-carrying radical is converted to one which is too unreactive to carry on the chain.

Autoinhibition. This action, sometimes called allylic termination, is exhibited by monomers which contain the highly reactive allylic C—H linkage. Free radicals are capable of hydrogen atom abstraction from copresent molecules as well as of addition to an unsaturated linkage. The ease with which this abstraction reaction is carried out by a given radical is a function of the reactivity of the C—H linkage which is attacked. The reactivity of radicals containing these C—H linkages increases in the order aryl < primary alkyl < secondary alkyl < tertiary alkyl < allyl < benzyl. Because of this high reactivity, a monomer containing an allylic C—H, allyl acetate, for example, functions as its own retarder, as shown by sequence (3). The resonance-stabilized allylic radical will react with the monomer only very slowly. The predominant further reaction is dimerization. Not only is polymerization slowed in this case, but also the molecular weight of the polymer formed in the reaction is low.



$$\begin{tikzcd}[column sep=tiny] & \sim M\cdot + CH_2{=}CHCH_2OAc \arrow[dl] \arrow[dr] & \\ \sim MCH_2CHCH_2OAc & & \sim MH + CH_2{=}CHCHOAc \end{tikzcd} \qquad (3)$$

Miscellaneous inhibitors. Oxygen, iodine, and nitric oxide interact rapidly with free radicals to yield stable products. Inclusion of these materials in polymerizing systems thus leads to effective inhibition. It is of particular interest that oxygen will copolymerize, under carefully controlled conditions, with certain monomers, styrene, for example, to yield high-molecular-weight polymeric, peroxides. The repeating unit is shown in formula (4). Iodine and nitric oxide have been used extensively to detect and in some cases to identify alkyl free radicals in nonchain as well as in chain reactions. This method has proved to be of great value in defining primary processes in photochemical decompositions of aldehydes and ketones. *See* POLYMERIZATION.

$$[\text{—}CH(C_6H_5)CH_2OO\text{—}]_n \qquad (4)$$

Inhibition of corrosion. Metallic corrosion in conducting media is electrochemical in nature. Local electrolytic cells are set up because of the presence of impurities, crystal lattice imperfections, or strains within the metal surface. The result is dissolution of the metal from the anodic regions. Corrosion inhibitors now in use may operate at the anodes or the cathodes or provide physical protection over the entire surface.

Anodic inhibitors. These are mild oxidants which reduce the open-circuit potential difference between local anodes and cathodes and increase the polarization of the former. Sodium chromate and sodium nitrite are most commonly used. The former is used in air conditioners, refrigeration systems, automobile radiators, power plant condensers, and similar equipment. Sodium nitrite finds special use in the protection of petroleum pipelines. It is effective even on rusty, mild steel. An extension of the nitrite type is the use of nitrite salts of secondary amines as vapor-phase inhibitors. The inclusion of a salt such as dicyclohexyl-ammonium nitrite with a packaged steel object provides effective protection against corrosion.

Cathodic inhibitors. Compounds such as calcium bicarbonate and sodium phosphate, in an aqueous medium, deposit on metal surfaces films that provide physical protection against corrosive attack.

Organic inhibitors. These are usually long-chain aliphatic acids and the soaps which are derived from them. Adsorption of these compounds on metal surfaces gives a hydrophobic film which protects the metal from corrosion by many agents. As little as 0.1% of palmitic acid, for example, is sufficient to protect mild steel from attack by nitric acid. *See* ANTIOXIDANT; CATALYSIS; CORROSION; FREE RADICAL.

Lee R. Mahoney

Bibliography. D. Braun et al., *Polymer Synthesis: Theory and Practice: Fundamental, Methods, Experiments*, 4th ed., 2004; V. S. Sastri, *Corrosion Inhibitors: Principles and Applications*, 1998.

Ink

A dispersion of a pigment or a solution of a dye in a vehicle (carrier), yielding a fluid, paste, or powder that can be applied to and dried on a substrate. Printing, writing, marking, and drawing inks are applied by several methods to paper, metal, plastic, wood, glass, fabric, or other substrates. Inks perform communicative, decorative, and even protective functions.

Printing Ink

Printing ink is by far the most widely used and technically developed type of ink.

Classification. Printing inks may be classified according to their characteristic physical properties (such as paste or liquid), the method of application, or other considerations such as the manner of drying. The composition of an ink can be oil-, solvent-, or water-based. The ink may be a high-viscosity paste, a low-viscosity liquid, or even a powder, as is the case for electrophotographic toner inks. The various drying manners are oxidation, evaporation, penetration, precipitation, polymerization, and radiation curing. Over 100,000 new ink formulations are created each year to meet the needs and requirements of printers.

The most important commercial printing methods that use inks are lithography, flexography, gravure, and letterpress. Economic factors as well as efficiency and speed of press preparation and setup are the key factors for their use. Lithography is currently the leading method of printing in the world and has been so for many decades, particularly in the publication and commercial fields. The next most widely used method is flexography, particularly in packaging printing, followed by gravure printing for both publication and packaging. Letterpress (nonflexographic) is declining in use and is now well behind the other three methods. Other printing methods include screen printing, inkjet, electrophotographic, stencil, and electrostatic. While inkjet and electrophotographic inks are widely used for computer hardcopy output, the quantities of ink consumed are still relatively low as compared to commercial graphic arts printing methods. *See* INKJET PRINTING; PRINTING.

Inks can be applied in four-color (screened) processes, as special spot colors (solid or screened), or as type or line work. Some end uses are newspapers, magazines, commercial, folding cartons, books, corrugated boxes, paper bags, wrappers, labels, metal containers, plastic containers, plastic films, foils, food inserts, sanitary papers, and textiles.

Formulation. Inks are composed of four major components: colorants, vehicles, additives, and solvents. Colorants (pigments, toners, and dyes) provide the desired color and contrast with the substrate. Vehicles, or varnishes, act as carriers for the colorants during the printing operation and, upon drying, bind the colorants to the substrate. Additives influence the printability, film characteristics, drying speed, and end-use properties. Solvents (including water) are used in the formation of the vehicles, and as diluents to reduce ink viscosity and adjust the drying rate. The ink components are weighed, mixed, and dispersed to form the finished product.

Colorants. These are used in the form of a dry powder, predispersed paste (called a flush), or liquid concentrate. Their selection critically influences the color, hue, permanency, opacity, gloss, rheology, and print quality. Thousands of colorants are available for matching the desired print color. Pigments containing cadmium, arsenic, mercury, antimony, lead, selenium, or significant impurities are considered undesirable for food packaging or children's products. *See* DYE; PIGMENT (MATERIAL); RHEOLOGY.

Vehicles. These are usually natural or synthetic resins or polymers dissolved in solvents, oils (hydrocarbon or vegetable), or liquid resins. Many ink resins are of moderate molecular weight to achieve good solubility. Synthetic resins with higher molecular weights have been developed for inks used on high-speed presses. *See* POLYMER; RESIN.

Inks solvents typically are emitted to the atmosphere and often are classified as volatile organic compounds and hazardous air pollutants. The primary solvent used in publication gravure inks is toluene because it will dissolve the high-molecular-weight resins used to make high-performance inks. However, publication gravure avoids solvent emissions by using an activated carbon system to recover the toluene. Water has become an important solvent for liquid flexographic and gravure printing inks that are used for printing porous substrates such as paper, particularly in packaging printing.

Drying methods. Vehicle selection or its composition is dictated by the ink drying process and the substrate and end-use requirements. The ink vehicles are best categorized according to their drying mechanism and application. Both physical and chemical reactions may take place during the solidification, or drying, step.

Evaporation. Solvent evaporation is accomplished by convective or radiant heat. An obstacle associated with converting to water-based inks is the drying process. Water has a higher latent heat of evaporation than organic solvents, resulting in slower drying and lower press speed. To alleviate this problem, a certain amount of organic solvent, such as an alcohol, is added to the ink. In addition, water-based inks require much higher color strength than solvent-based inks in order to facilitate drying by reducing the wet-ink film thickness.

Letterpress and lithographic heatset inks are based on vehicles consisting of modified rosin and hydrocarbon-type resins. These are dissolved mainly in aliphatic hydrocarbons. Heatset vehicles are often gelled for faster and sharper tone printing. *See* ROSIN.

Flexographic solvent-based inks use one or more synthetic polymers dissolved in lower alcohols, often in combination with small amounts ($<10\%$) of esters, hydrocarbons, or glycol ethers. Emulsions or microdispersions of acrylic copolymers, polyamide resins, and polyvinyl copolymers, along with nitrocellulose, are used in flexographic inks for printing films and foils. Since flexographic printing presses may use rubber plates and rollers, the tolerance for many solvents is often below 10%. *See* ALCOHOL; COPOLYMER; ESTER; HYDROCARBON; POLYACRYLATE RESIN; POLYAMIDE RESINS; POLYVINYL RESINS.

Gravure inks for packaging applications contain a variety of polymers dissolved in blends of organic solvents. The use of these solvents has been reduced, or even eliminated in many cases, since they contributed significantly to the volume of solvent emissions. One goal has been to replace most or all the solvent content with water. However, water-based inks have not been developed with all of the physical properties required by many packaging applications, especially resistance to water and heat.

Gravure systems for publication printing use calcium or zinc salts of rosins, rosin esters, and phenolic modified rosin or hydrocarbon resins dissolved in aliphatic or aromatic hydrocarbon solvents.

Precipitation. Binder (resin and oil) precipitation occurs when water is absorbed by a hygroscopic solvent. Steam or moisture from the air causes the binder to separate from the solvent; the solvent then either penetrates a porous substrate or evaporates. Typical binders used in letterpress and dry-offset printing inks are fumarated rosin and occasionally a maleated drying oil.

Oxidation. This is the reaction of atmospheric oxygen with unsaturated fatty acids present in vegetable oils or in their alkyds. Cobalt and manganese salts of fatty and aromatic acids act as catalysts for the reaction. Letterpress and lithographic sheetfed inks dry (solidify) by this mechanism. *See* DRYING OIL.

Thermal polymerization. This reaction normally proceeds via a polyol-amine condensation path or through the Diels-Alder diene-addition mechanism. Lithographic, dry-offset, letterpress, and screen-printing inks for metal, paper, and glass frequently use this method. Alkyds or polyesters with free hydroxyl groups condense with melamine resins in the presence of acid catalysts. Alkyds or oils with diene-conjugated unsaturation polymerize via the addition and oxidation mechanism simultaneously. *See* AMINE; DIELS-ALDER REACTION; POLYESTER RESINS; POLYMERIZATION; POLYOL.

Gel-fusion. Gel-fusion takes place when the binder, composed of a nonsolvent plasticizer and a dispersed polymer, is heated to the temperature at which polymer solubility is achieved, and upon cooling a solid ink film results. Letterpress, dry-offset, and some lithographic inks contain a poly(vinyl chloride) resin dispersed in a suitable plasticizer.

Quick setting. This method relies on a two-step mechanism involving the critical solubility of a resinous binder and the porosity of the substrate. A vehicle with borderline solubility releases part of its solvent to the substrate, precipitating the resins, and becomes tack-free. The second drying step proceeds via the conventional oxidation process. Lithographic and letterpress paper and folding-carton inks for sheetfed presses use this drying principle.

Absorption. Absorption setting is achieved through the penetration (sponge effect) of the total vehicle or its liquid portion into a porous substrate. Newspaper inks, based on mineral (hydrocarbon) oil or on rosin, or gilsonite (variety of asphalt) in mineral oil varnishes, set, although never fully dry (**Fig. 1**).

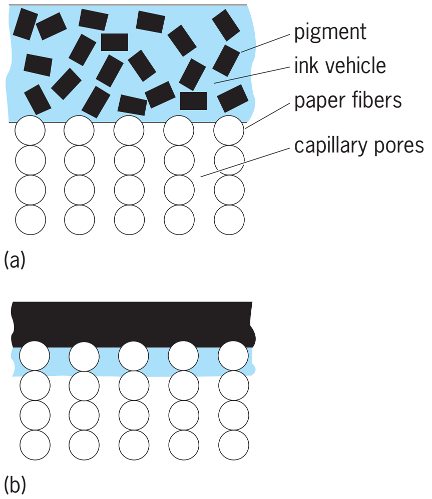


Fig. 1. Ink drying by absorption setting. (*a*) Ink film on uncoated paper immediately after printing. (*b*) Ink film after absorption of the vehicle by the paper.

Cold setting. In this method, inks are applied in a molten state and solidify on cooling. Cold-setting inks (also called hot-melt inks) contain no solvents. Instead, they are based on a combination of waxes (petroleum and synthetic) and colorants. *See* WAX, PETROLEUM.

Radiation curing. This method uses ultraviolet (UV) or electron-beam radiation to cure (dry) the ink. The principal advantages of this method are lower energy consumption, instant curing, low-temperature curing, and reduced solvent emissions.

Ultraviolet light, generated by medium- or high-pressure mercury lamps, activates photoinitiators to generate reactive species, called free radicals, which initiate the polymerization of acrylic monomers and oligomers in the vehicle. Market growth of UV inks has been mainly in printing sheetfed litho cartons and plastic cups, and in metal decorating. UV curing of overprint coatings, which impart gloss and abrasion resistance to book covers, magazines, and packaging materials, is very popular and growing in use. UV curing is also widely used in narrow-web flexographic printing of labels and in screen printing. *See* PHOTOCHEMISTRY; ULTRAVIOLET RADIATION.

Electron-beam inks are chemically similar to ultraviolet inks except that they do not require photoinitiators to polymerize acrylic monomers and oligomers since high-energy electrons are able generate free radicals. *See* CHARGED PARTICLE BEAMS.

Infrared drying. Infrared heating is occasionally used to help set lithographic inks. Although infrared heating is not used with web offset presses, it finds some utility on sheetfed litho presses. Inks containing some volatile solvents are required. *See* INFRARED RADIATION.

Microwave drying. This method is used occasionally in liquid inks that contain water as the main solvent, but finds very limited use because of poor energy efficiency in creating heat. *See* MICROWAVE.

Additives. A simple suspension of pigment in varnish or solvent usually does not produce a commercially satisfactory ink. Many other materials must be added in order to provide good performance. These additives include plasticizers, wetting agents, waxes, shortening compounds, reducers, thickening agents, antiskinning agents, and antipinhole compounds.

Plasticizers are commonly used in flexographic printing inks. They make flexographic inks softer, more flexible, and more adherent to the substrate. Resins are often rigid and brittle, and adhere poorly to flexible substrates. Selection of the plasticizer depends on the printed product and the resin in the ink. Since the boiling points of plasticizers are generally high, they do not evaporate in the dryer but become a permanent part of the ink film. Plasticizers include various phosphates, sulfonamides, chlorinated materials, citrates, adipates, polyglycol derivatives, resinates, and phthalates.

Most printing inks contain waxes to improve rub and scuff resistance. The waxes used in inks are often a compound comprising a fine dispersion of several waxes in a vehicle. They may also be a micronized (very finely divided) dry powder. Common classifications of waxes include animal (beeswax, lanolin), vegetable (carnauba, candellila), mineral (paraffin, microcrystalline), and synthetic [polyethylene, poly(ethylene glycols), polytetrafluoroethylene (PTFE or Teflon)]. A wax that is too soluble in an ink will come out of solution when the ink cures, form a waxy film on the surface, and prevent wet ink from trapping (overprinting) on dry ink films. This problem is commonly referred to as crystallization. Excessive use of wax should be avoided to prevent problems associated with ink piling. Too much wax in the ink may cause increased drying time, trapping problems, and decreased gloss.

Wetting agents promote the dispersion of pigments in ink varnishes. Wetting agents for lithographic inks must be carefully selected; otherwise, they can cause excessive emulsification of the dampening solution into the ink and other associated problems. Wetting agents may also reduce the surface tension of aqueous flexographic and gravure inks in order to improve their wetting properties on plastic and foil substrates.

Shortening or antimisting agents reduce ink flying or misting of ink at high speeds. The addition of a wax or an organo-clay compound shortens an ink and can reduce misting. Improper amounts of antimisting agents can interfere with ink flow on the press and aggravate ink piling.

Reducers, such as heatset oils or other petroleum solvents, are occasionally added to soften and reduce the tack of a litho ink. A thin litho varnish, grease, boiled linseed oil, or a light-oil-modified alkyd will also reduce the tack of an ink.

Thickening agents are used to modify the viscosity and tack of paste inks to adjust them to specification.

Antiskinning agents are antioxidants that slow the drying of sheetfed offset inks, so that they will not skin (dry) in the can. If they are sufficiently volatile, they will not greatly retard the drying of the printed ink.

Defoamers and antifoamers for aqueous flexo or gravure inks are blends of surface-active organic liquids, surfactants, metal soaps, hydrophobic silica, and other ingredients, with or without silicone modification. They may be used during the preparation of aqueous inks to prevent foam buildup, and may be added during use for the same purpose.

Solvents. Solvents dissolve the oils, resins, and additives used to produce the vehicles that carry the pigment. A wide range of solvents are used in printing inks, including hydrocarbons, heatset oils, alcohols, glycol ethers, ketones, and esters. Polar solvents such as alcohol, ethers, ketones, and esters are useful in flexographic and gravure vehicles for dissolving polar resins such as shellac, cellulose esters, phenolics, and acrylics. The hydrocarbons, which are nonpolar solvents, are suitable in lithographic and letterpress vehicles for dissolving compounds such as drying oils and rosin-modified phenolics and maleics.

Two classes of hydrocarbons are used: aliphatic and aromatic. Aromatic hydrocarbons are more powerful resin solvents than aliphatic hydrocarbons. Toluene is used extensively in publication gravure inks. Mineral spirits, naphtha, and kerosene consist principally of aliphatic hydrocarbons. These solvents are used in inks for lithographic and letterpress printing. In flexographic inks, aliphatic hydrocarbons are primarily used as cosolvents in polyamide formulations or as inexpensive diluents. Aromatic hydrocarbons are used sparingly in flexographic inks because they tend to swell rubber and plastics.

Carefully fractionated (narrow-boiling) hydrocarbon heatset oils are solvents for most of the resins found in lithographic and letterpress inks. The choice of solvent with the proper boiling range depends on the type of ink. The boiling points of aliphatic hydrocarbon solvents for heatset inks vary from about 400 to 600°F (200 to 300°C), but each designated fraction boils over a much narrower range. The choice of heatset oil depends principally upon the type of resin used and the printing conditions for which the ink is intended. The lower the boiling point, the lower the temperature required to dry the ink. Selection of the solvent depends not only on the boiling point but also on the characteristics of the dryer, press, press speed, ink coverage, and paper stock.

Alcohols are used extensively in flexographic and packaging gravure inks and in some screen-printing inks. Their dissolving power, evaporation range, and mild effect on rubber make them important flexographic ink solvents. Alcohols include ethyl, normal propyl, and isopropyl alcohols. Ethyl alcohol and isopropyl alcohol in flexographic inks are used primarily in polyamide-based vehicles and as additives in water-based formulations. Normal propyl alcohol is used to slow down the evaporation rate of inks containing alcohols that are more volatile.

Glycol ethers are alcohols with ether linkages and are excellent solvents for most synthetic resins. The glycol ethers used in printing inks are soluble in both water and alcohols. Glycol ethers are used in flexographic and packaging gravure inks, as well as screen printing inks. They are good solvents for nitrocellulose and many other resins, including shellac, vinyl acetate, and rosin derivatives.

Ketones are strong solvents. Acetone is used in nitrocellulose lacquers and packaging gravure inks. It is miscible with water, alcohols, ethers, oils, hydrocarbons, halogenated hydrocarbons, fatty acid esters, and most organic solvents. Methyl ethyl ketone and methyl isobutyl ketone are used in gravure and screen printing inks. Cyclohexanone and isophorone are used in some screen-printing inks. Ketones are rarely used in flexographic printing inks because they swell rubber and plastic plates.

Esters are generally good solvents for nitrocellulose. The rapidly drying solvents ethyl and propyl acetates are used extensively as cosolvents in flexographic and packaging gravure inks. *See* SOLVENT.

Testing. Inks are subjected to numerous laboratory tests during and after their manufacture. The most common properties that are checked are fineness of grind, color and strength, gloss, abrasion resistance, drying speed, and rheology. Other characteristics important for special types of inks are tack, misting, water emulsification, transferability, and overprint (trap) quality for lithographic inks; adhesion, odor, and solvent retention for flexographic or gravure laminating inks; ice water, oil, and soap sensitivity for surface-printed packaging inks.

Fineness of grind. The most precise method for testing fineness of grind is the actual measurement of the pigment particle size, the size range, and its population. Several approaches are available to test the particle size and distribution, which indirectly influence an ink's color strength, printability, gloss, and rheology. Many light-scattering instruments are now available to determine the particle size distribution of inks, but they are expensive and time-consuming and used mainly in research and development. Particle size and distribution can also be determined by optical and electron microscopy. For routine quality-control tests, a fineness of grind test for oversize particles is done on a grindometer (**Fig. 2**). *See* LIGHT-SCATTERING TECHNIQUES.

Color. This is one of the key properties of an ink because of its functional and esthetic value. There are several methods for controlling the printed color, which depends on both the thickness of the ink film and its intrinsic color. Reflection densitometers are used in conjunction with color filters to measure print optical density at press side, which is proportional to printed ink film thickness. The actual printed ink color can be measured with a portable computerized spectrophotometer, which yields absolute values for hue, lightness, and chromaticity. These data are often specified as a quality-control parameter by the printer for critical jobs.

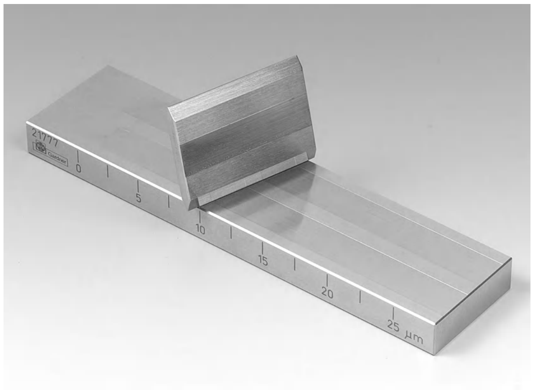


Fig. 2. Grindometer.

Fig. 3. Brookfield digital viscometer. (*BYK-Gardner USA*)

Rheology. Ink rheology influences press transfer, print fidelity, overprinting, and gloss characteristics. Several instruments are available to measure viscosity, yield value, and flow characteristics (**Fig. 3**). Special instruments have been designed for dynamic high- speed press applications, where determination of the viscous and elastic components in ink rheology has become more critical. Viscometers usually are interfaced with a computer to handle the calculations, program the tests, and analyze the output. *See* MICROCOMPUTER; RHEOLOGY.

Tack. This measure of film splitting (separation) force applies to lithographic and letterpress paste inks. It controls the transfer of ink from plate or rubber blanket to substrate and influences the overprinting characteristics. The tack is measured with roller distribution instruments known as inkometers and tackoscopes (**Fig. 4**). Higher press speeds require lower ink tacks to prevent tearing, or picking up lint from, the paper.

Transfer. The laboratory determination of ink transfer and printability is normally done by print proofing. This test is used to predict the color of an ink

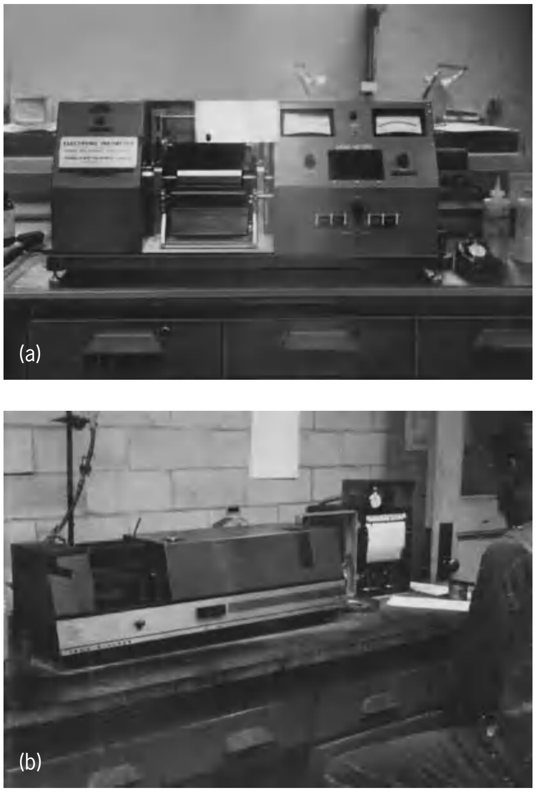


Fig. 4. Instruments for the measurement of ink tack. (a) Inkometer. (b) Precision tack tester. (*Prüfbau Instruments*)

applied to a specific substrate or to a predeposited ink film.

For lithographic inks, proofs can be made with a simple press, where a hand-cranked impression roller transfers ink from a plate to a substrate. For accurate high-speed tests, automatic proof presses are used (**Fig. 5**). Here both ink and fountain solution are applied in a controlled amount to a substrate.

For flexographic inks, a manually operated tool consisting of an engraved (anilox) cylinder and a rubber roller is used. The ink is metered through the nip between the cylinder and the roller. The substrate is put on a flat resilient surface, and the ink is applied at hand speed. Several motor-driven flexographic proofers are also available, which provide more consistent and accurate prints.

For gravure inks, small gravure presses are used. Here the engraved cylinder is partially immersed in an ink fountain. The excess ink is wiped off the cylinder by a metal blade and printed onto a substrate in web or sheet form.

Ink-water balance. In lithographic printing, the most important press settings involve the ratio of the ink feed rate to the fountain solution feed rate. This is called the ink-water balance and is critical to good printing. Since a lithographic plate has a flat printing surface, the image must be chemically differentiated by keeping the image areas wetted with ink and the nonimage areas wetted with water at all times. At present, there is no automatic control for balancing the two feed rates, even though the ink feed can be controlled to achieve a given optical density. The press operator's skill in setting the fountain solution feed is critical to good printing.

Computer color control. Several computer color-matching systems are used in the ink industry for replicating ink (color) formulas, matching new colors, reducing costs, and quality control of pigments. They automate color matching mathematically by blending stored spectral curves of pigments of different color. In these systems, a spectrophotometer is coupled to a computer to produce very rapid results.

Manufacturing processes. Production of most inks is predominantly a batch process, often consisting of wetting, mixing, milling (dispersion), and filtering steps. It must be efficient and flexible, capable of producing a consistent and high-quality product. Some larger-volume inks (such as newspaper, web offset, and publication gravure inks) are mass-produced in a semicontinuous mode. Such lines are automatically controlled, employing microprocessors and process sensors.

The three main steps in ink making are mixing, dispersion, and filtering. Mixing and its intensity differ with the characteristics of various inks and depend on the mixer speed, size, and configuration. Mixing speeds range 100–5000 revolutions per minute, which requires proper sizing of the motor for adequate energy input. Dispersion reduces agglomerated pigment particles and even crushes larger pigment crystals. This step controls the color strength (and value) and influences ink rheology.

Dispersion of paste (letterpress and offset) inks normally is done with three-roller mills or media mills (**Fig. 6**). Liquid (flexographic and gravure) inks are dispersed with closed-head media mills and with ball mills. Colloid mills are used occasionally for

Fig. 5. Multipurpose printability tester. (*Prüfbau Instruments*)

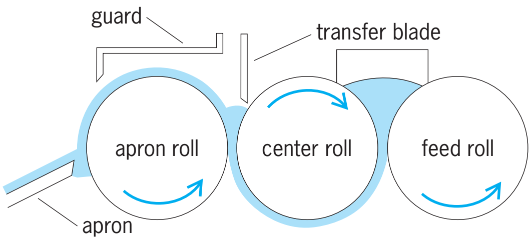


Fig. 6. Three-roll mill.

newpaper inks and some liquid inks. *See* GRINDING MILL.

Filtering is done with special cloth bags for liquid inks and vehicles, and with bags or cartridges for paste inks. This step removes impurities, agglomerated pigment, fibers, and grit. Media milling and filtering are often used in the continuous production of large-volume process color inks. Magnetic filters are also used for gravure inks.

Ink preparation. Typically, an ink can be prepared by either of two processes. In the dry grind process, all the ingredients (such as varnish, pigment, compounds, talc, oil, and solvent) are placed in a mixer. After mixing and dispersion, the ink is tested and compared to a standard for tack, viscosity, grind dispersion, color, strength, and drying. The ink is adjusted as needed and then packaged. The ink is distributed in sizes ranging from small cans to commercial-size reusable tote bins or tank trucks.

In the base process, the pigment is mixed with only that part of the varnish that imparts the proper viscosity to fit the equipment being used, forming a concentrated base. The base is then mixed with the other ingredients, tested, and packaged. For certain lithographic inks, the base is not made from dry pigments but from pigment presscake. Here the water-wet presscake is mixed with varnish in a heavy-duty dough mixer; the water is flushed out via the shearing action and a flushed base is produced. Another method for flexographic or gravure inks is the use of color chips. The dry pigment is mixed with resin plus some plasticizer and milled on a two-roll mill. The resulting dispersion is dissolved in solvent, and other ingredients are added to make the final inks. Liquid pigment dispersions for the preparation of finished inks are also available. They generally are prepared from pigment, solvent, resin, and a dispersant. They are diluted with a vehicle to prepare an ink.

Special Inks

A number of special-purpose inks have been developed; these may be used for complex applications other than simple printing.

Laminating inks. Technically sophisticated formulations are required for producing laminating inks for flexible film packaging. Unlike other applications, where the ink is applied onto the outside surface of a substrate, a laminating ink is applied via flexography or gravure onto the reverse side of a thin film, which in turn is laminated to other films, foils, or papers. Such inks and structures are used for packaging snack foods, beverages, and medicines. Special ink formulations based on synthetic resins are used for more demanding products such as aseptic food packages. *See* FILM (CHEMISTRY).

Electrophotographic inks. The digital color imaging process technology that led to the development of on-demand printing requires different types of ink than those used in conventional printing. Two special inks have been developed: dry toners and liquid toners.

In these systems, images and characters are generated and stored in computers and then transferred via digitally driven laser beams onto photosensitive drums, resulting in a latent image of the desired object. Dry or liquid toner inks are then fed onto the surface of the drum. Next, electrostatic forces within the latent image on the drum attract the toner, as it possesses an opposite charge. The inks are transferred from the drum to an appropriate substrate and fixed by heating. In some instances, an intermediate tool (offset blanket) is used in the transfer process between the drum and the substrate.

Electrophotographic inks must be capable of carrying the electrical charge required by the specific type of drum in a given printing system. Dry toner inks are composed of fine resinous thermoplastic powders, and liquid toners are composed of finer powders dispersed in an inert dielectric carrier such as an isoparaffinic hydrocarbon. *See* PHOTOCOPYING PROCESSES.

Engraving (intaglio) inks. These are inks designed for high-quality and security printing, as required for paper currency, stock certificates, and formal letterheads. The image is engraved below the plate surface in multiple lines on the steel plate. A thick layer of ink is applied to the plate, filling the recesses of the engraving, and the excess ink is removed by wiping with paper or a soft plastic-covered roll and a special washing fluid. The rheology of the ink is such as to allow for easy wiping (a property called short body). In addition to colored pigment, a large amount of colorless extender is used to assure that the thick film of ink applied will wipe properly from the nonimage area (plate surface). In the case of paper currency, special ingredients are used to assure such properties as durability, stiffness to accommodate money-changing machines, magnetic properties, launderability, and chemical and abrasion resistance.

Transfer inks. These inks may be printed by the flexographic, gravure, or lithographic process onto a temporary substrate (usually paper) for later transfer onto the final objects they are intended to decorate. Heat-transfer inks that contain sublimable dyes are printed on paper and are transferred to other substrates using heat and pressure. The most popular application is in imaging of fabrics, especially polyester. Decalcomania inks are normally transferred from paper onto objects such as ceramics, windows, or outdoor signs. *See* TEXTILE PRINTING.

Waterless lithography. Waterless offset inks contain many of the same ingredients found in conventional lithographic inks. The difference is that waterless inks have vehicles that produce higher initial viscosities than conventional inks. Waterless offset inks also may have different tack than conventional offset inks in order to increase their cohesion. The viscosity and tack differences of these inks relate to differences of the printing plates used. Conventional offset lithographic plates have a hydrophilic nonimage area that attracts water or dampening solution, and an oleophilic image area that repels water and attracts oily ink. Waterless offset plates use a silicone polymer for nonimage areas of the plate and a metal or ink-accepting image area. The silicone polymer's low surface energy resists ink, as long as the ink's viscosity is high enough for the ink to be more cohesively attracted to itself than to the silicone.

The viscosity of ink decreases rapidly with increasing temperature (about 10% per degree Celsius). In conventional lithography, the presence of water in the press's dampening solution cools the ink, allowing it to maintain a more stable viscosity. The viscosity of waterless offset inks is maintained on the press by a temperature control system, either a plate cylinder cooling system or an ink oscillator (vibrator) cooling system. Waterless offset inks are formulated at varied viscosities for different press speeds.

Duplicating inks. Essentially, these are modified sheetfed lithographic inks for special small presses used by offices and print shops. Duplicating printing presses are characterized by two different and relatively simple ink-distribution roller arrangements. In the integrated ink-water system, a single roller is used to apply both ink and fountain solution to the plate. As a result, the ink must be capable of releasing the water easily under modest contact pressure. The ink formula is similar to other lithographic inks, except that it is formulated to be fast-setting without heat.

Other inks. Inks developed for ballpoint pens are thin fluids of high color tint (pigment or dye) strength. They usually contain dye solutions or fine pigment dispersions in various low-viscosity vehicles. The inks must be free of particles and resist premature drying to feed to the paper without clogging. Rapid penetration into the paper accomplishes drying.

Stamp-pad inks are supplied in cloth or foam rubber pads and transferred by pressure to raised rubber type or image, which is then stamped or pressed against the substrate. The inks must remain nondrying on the pad, yet rapidly penetrate into the stamped substrate. Dyes are often dissolved in glycols or other low-volatility solvents to form these inks.

Regulations. In the United States, federal and local governments regulate ink manufacture and use. The most important regulation affecting the industry is the Clean Air Amendment of 1990, administered by the Environmental Protection Agency (EPA). This regulation monitors emission of volatile organic compounds and hazardous air pollutants. In addition, the industry must conform to two complex federal government regulations: the Resource Conservation and Recovery Act (RCRA) and the Comprehensive Environmental, Response, Compensation, and Liability Act (CERCLA), including its self-standing Superfund Amendments and Reauthorization Act (SARA). These laws, also administered by EPA, have led to major changes in the industry such as avoidance of pigments containing heavy metals, development of inks capable of easy deinking for recycling of news and office paper, and recycling of ink waste. In the more complex area of toxicology, the ink industry is regulated by the Toxic Substance Control Act (TSCA).

Worker safety is ensured by the Occupational Safety and Health Act (OSHA), administered by the Department of Labor. This regulation is mainly concerned with the vapors (solvents) and liquids (resins) with which the ink manufacturer or the printer comes into contact during routine work. *See* ENVIRONMENTAL ENGINEERING; INDUSTRIAL HEALTH AND SAFETY. Robert W. Bassemir

Bibliography. T. M. Destree (ed.), *Lithographers Manual*, 9th rev. ed., GATF Press, 1994; N. R. Eldred, *What the Printer Should Know about Ink*, 3d ed., GATF Press, 2000; N. R. Eldred, *Chemistry for the Graphic Arts*, 3d ed., GATF Press, 2001; G. G. Field, *Color and Its Reproduction*, 2d ed., GATF Press, 1999 Flexographic Technical Association, *Flexography: Principles and Practices*, 5th ed., FFTA, 1999; R. H. Leach (ed.), *The Printing Ink Manual*, 5th ed., Blueprint-Chapman & Hall, 1993; J. MacPhee, *Fundamentals of Lithographic Printing*, vol. 1, GATF Press, 1998.

Inkjet printing

A method of forming a hardcopy image (text or graphics) by generating very small drops of ink and precisely depositing them on a substrate. This concept was introduced commercially in the mid-1980s to replace dot-matrix (low cost but slow and noisy) and laser-based (high speed and high resolution but also high hardware cost) printing systems for home and office applications. As inkjet color inks and photographic-quality paper development progressed, a new goal of supplanting traditional silver halide photographs was established. Today, the home photographer can produce inkjet prints that exceed the quality and light-fade stability of color photographic prints. *See* PRINTING.

Inkjet print quality has improved continuously by using a systems approach of providing smaller drop volumes along with better media (paper) and inks. The colorants used in the inks may be dyes (water/solvent-soluble) or pigments (insoluble). Dye-based inks historically provided more vivid color and were somewhat easier to manufacture and keep in suspension, but pigment-based inks are closing the gaps in these areas. Pigmented inks, which do not penetrate the media as readily as dye-based inks, have been developed with the equivalent quality of laser printers or typewriters for black text printed on plain paper. Compared to dye-based inks, pigmented inks provide water-fastness, light-fade resistance, and better smear resistance. Recently, color pigmented inks have found use in large format inkjet printers for outdoor signs. Photo imaging and color adoption in the office applications have been instrumental

Comparison of inkjet printhead technologies

		Drop-on-demand	
	Continuous (piezo)	Thermal (TIJ)	Piezo (PIJ)
Advantages	High volume High frequency Ink flexibility Long throw distance Many substrates	High efficiency Low cost Integrated chip processes Integrated drive High frequency Nozzle density	High efficiency Ink flexibility Long throw distance Many substrates Multiple drop volumes Long-life head
Disadvantages	Nozzle density High head cost	Ink flexibility Unwanted heat Head reliability	Nozzle density High head cost Electrical integration

in increasing use of inkjet products. *See* DYE; INK; PAPER; PIGMENT (MATERIAL).

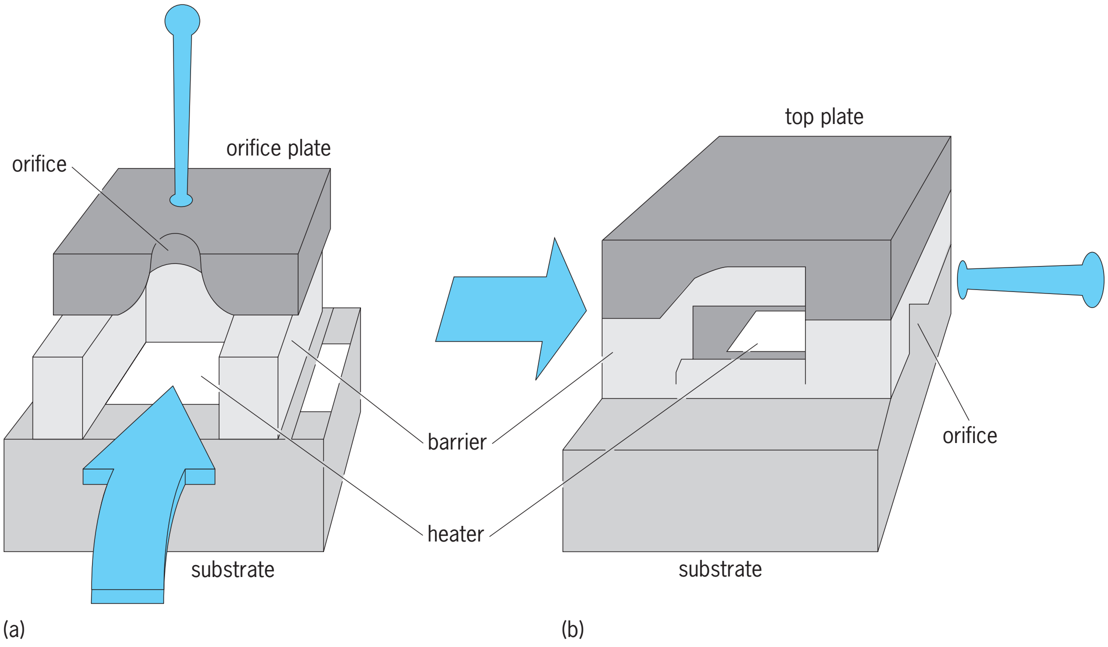
The human eye can detect dots on paper 25 micrometers (0.001 in.) or smaller if the color is dark (high optical density). A dot of this size on photo paper requires an ink drop of 2 picoliters or less, assuming typical dot spreading on the media. Therefore, drop volume reductions beyond 1 pL are probably not necessary for grain-free photos. Color printing speed and quality also have been improved by adding more colors, some with reduced dye or pigment loads of the black, yellow, cyan, and magenta colorants, but this adds product cost. In the future, important product features will be speed and ease of use.

Printhead technologies. There are two popular techniques used to jet inks with adequate drop ejection velocity, size, and directional control. They are continuous and drop-on-demand (DOD), as seen in the **table**. The earliest inkjet (continuous) used a piezoelectric material that changes shape when a voltage is applied. This shape change is used to increase or decrease the volume of ink confined in the ink chamber. A column of ink exits the firing chamber through a small hole or nozzle and then breaks into a droplet. The ink flow is continuous as a periodic voltage is supplied to the piezoelectric material. Unwanted drops are deflected into an ink-recycling channel. Piezoelectric materials are also used in DOD printheads. Thermal inkjet (TIJ) is another DOD technology and has gained over 75% of the market. Here the ink is superheated, and the resulting large volumetric expansion is used to propel the droplet. Typical drop volumes range from 2 to 50 pL, with velocities of 10 m/s, and a head-to-media spacing of 1 mm. *See* PIEZOELECTRICITY.

Continuous inkjet. This technology is used primarily for high-volume applications such as magazine labels, can-dating codes, bar-code labels, and carton marking. Inkjetting frequencies of 100 kHz or more can be attained since the head is operated in a stable mode. The ability to switch between ink systems (such as dye and pigmented inks), known as ink flexibility, and the short distance from the head to the substrate are advantages. However, the nozzle packing density (number of nozzles per printhead area) and the high printhead cost have prevented this technology from penetrating consumer printing applications. New applications for continuous inkjet are emerging as new inks with improved durability and specific attributes that allow direct printing on glass, ceramics, fabrics, and other substrates are developed. Systems using continuous inkjet technology typically cost from $50,000 to nearly $1 million.

Drop-on-demand inkjet. In digital printing, the most efficient use of energy is to place ink only when and where needed, or DOD. The volume, velocity, and direction of each ejected drop are critical for maintaining consistent quality and operating conditions. A computer generates the desired image and sends the dot firing sequence to the printer, which either energizes a piezo crystal or fires a thermal resistor in a printhead with hundreds or thousands of small nozzles lined up in several rows. The printhead scans over the media, printing a swath of one or more colors. This process is repeated until the entire document is printed.

Thermal DOD inkjet. Since the mid-1980s, thermal inkjet has shown continuous improvement in quality, speed, and cost. A major breakthrough in the technology was the development of the replaceable printhead. This allowed both the ink and printhead to be placed in one inexpensive cartridge. There are two primary technologies used in thermal inkjet: edge and top shooter (see **illus.**). Both techniques use the physical principle of converting the electrical energy delivered to the firing resistor into heat in order to superheat a small volume of ink. Most thermal inkjet now use the top shooter design since it is more thermally efficient. Thermal inkjet primarily uses established integrated circuit manufacturing processes and equipment, so costs are generally lower than piezoelectric inkjet (PIJ). Its silicon-based substrate allows the integration of electronic drive circuitry, which reduces the overall printing system costs and improves performance. The latest thermal inkjet technology uses silicon etching for ink feed slots, thermal inkjet resistor formation, and ink channel/nozzle formation. Drop ejection frequencies of up to 36 kHz have been achieved. The upper limits for firing frequency are determined by the overheating and fluid refill characteristics of the chamber. Firing chamber densities are currently at 600 dots per inch (dpi). For most applications, acceptable text



Thermal inkjet printhead designs: (*a*) top shooter and (*b*) edge shooter.

and graphics quality can be achieved at 600 dpi, so a single pass of the printhead is sufficient. Higher-quality printing is achieved with several passes of the printhead at a sacrifice in printing speed. Thermal inkjet is lower-cost, and allows a higher degree of circuit integration and higher nozzle packing density than piezoelectric inkjet.

Piezoelectric DOD inkjet. Piezoelectric inkjet did not find wide acceptance until the mid-1990s, when costs, image quality, and ease of use began to match thermal inkjet. Early DOD piezoelectric inkjet piezoceramic printheads required very costly machining operations. The development of low-cost thick-film piezo elements and high-volume manufacturing contributed to this technology. Materials used in piezoelectric inkjet printhead construction are more robust than thermal inkjet, which has enabled them to jet a wider range of inks. In contrast to thermal inkjet, the piezoelectric inkjet printhead is permanent and only the ink cartridge is changed. One popular use is a phase change ink, or wax-based ink. The printhead is operated above the melting temperature of the wax, so the jetted drops "freeze" on the media. This allows a wide choice of substrates, but the printhead and maintenance costs are high. These products are used in the office and low-volume commercial markets. Since the ink is "squeezed" from the chamber, more control of the fluid dynamics is possible. As many as three different drop volumes ranging in an order of magnitude are possible from a single firing chamber, reducing thermal inkjet's advantage of higher nozzle packing density. Drop dynamics are similar to thermal inkjet with fewer unwanted satellite drops formed, but slightly lower drop velocity and operating frequency. Rob Beeson

Bibliography. R. Beeson, Is print quality an issue anymore—Or is it performance, performance?, *GIGA 25th Global Ink Jet and Thermal Printing Conference*, 2002; H. Le, Fabrication materials and process on ink jet printheads, *IS&T Non-impact Printing Conference Tutorial*, 2000; R. Mills, Ink jet in a competitive market place, *IMI 10th Annual European Ink Jet Printing Conference*, 2002.

Inland waterways transportation

The movement of cargo on inland and intracoastal waterways by barge. Barges move over 800 million tons of raw materials and finished goods each year just in the network of inland and coastal waterways of the United States. This waterways network is a "super water highway" that is an important part of the intermodal transportation system. Barges directly serve 87% of all major United States cities, accounting for 78% of all domestic waterborne traffic.

In the United States, barges move on the 25,543 mi (41,107 km) network of commercially navigable inland and intracoastal waterways (**Fig. 1**). Generally, barges are either pushed or towed by towboats on the inland waters. Barges require standard operating depths of at least 9 ft (about 3 m). Such depths are provided in 15,675 mi (25,226 km) of channel; the other 9868 mi (15,881 km) are less than 9 ft deep. Certain sections of the inland waterways are deep enough to support oceangoing vessels, including the Hudson River to Albany, New York; the Lower Mississippi north to Baton Rouge, Louisiana; the Houston Ship Channel; the Delaware River; the James River; and portions of the Columbia-Snake River.

Vessels. Four major types of barges are used on the inland and intracoastal waterways: the open hopper, the covered dry cargo, the tank barge, and the deck barge.

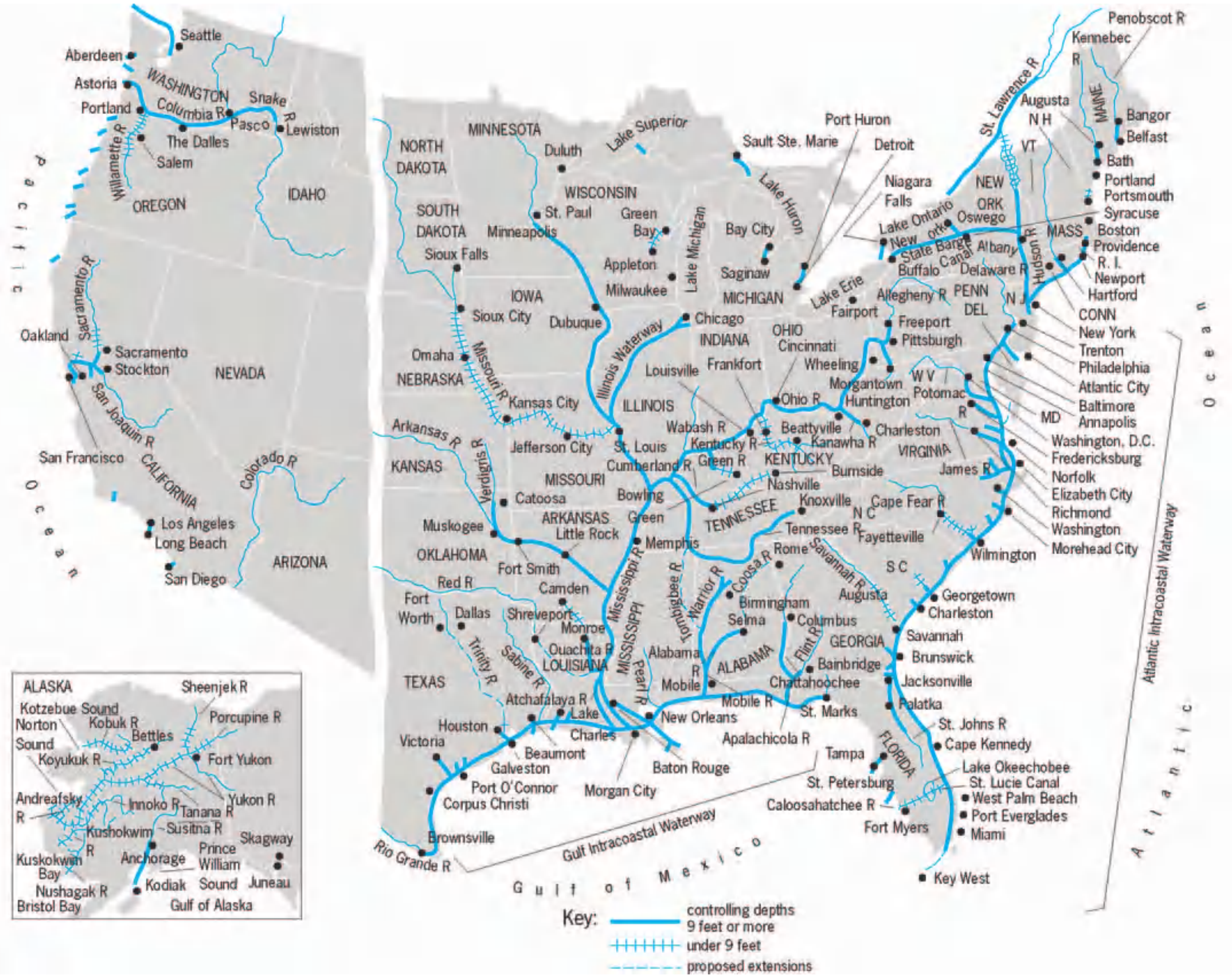


Fig. 1. Commercially navigable inland and intracoastal waterways of the United States.

The open hopper barge is used to transport cargo that does not need to be protected from the elements, such as coal, sand, and gravel. The capacity of this type of barge ranges from 1000 to 1300 tons (900 to 1200 metric tons), and the range in length from 175 to 290 ft (53 to 88 m).

Cargoes that need protection, such as grain, are shipped on covered dry cargo barges, which have 1000-1500-ton (900–1350-metric-ton) capacities and range in length from 175 to 200 ft (53 to 61 m). The covers of these barges may be the lift-on, lift-off type, or have a telescoping design, where the covers roll fore and aft on tracks installed on the barge.

Tank barges carry liquid commodities, such as petroleum and chemicals. The capacity range is 1000-3000 tons (900-2700 metric tons), and lengths vary from 175 to 290 ft (53 to 88 m). Many tank barges are divided into compartments—some heated to high temperatures, some below freezing—so that a variety of liquid cargoes can be carried in one tow.

Deck barges have flat tops, conducive to carrying containers, construction and drilling equipment, oversized equipment, and machinery, such as turbines for power plants. Deck barges are even used to carry components of the space shuttle from their construction site to the launch pad area. The roll on/roll off, or RO/RO, barge (**Fig. 2**) is a deck barge with multiple decks that can accommodate enormous numbers of containers, trailers, or rail cars, which can be wheeled directly onto or off the barge.

Over 27,000 of these non-self-propelled barges ply the commercial navigation system, offering a combined capacity of 5×10^7 tons (4.5×10^7 metric tons). These vessels receive propulsion power from approximately 4000 towboats and tugboats.

Towboats with barge flotillas commonly operate on the inland waterways, where they are fully protected by land on either side. The barges are kept together by wire rigging and are lashed against the towing knees of the towboat that pushes them ahead

(**Fig. 3**). The flotillas vary in size depending on the waterway on which they are operating. On the Upper Mississippi River above St. Louis, and on other waterways that are controlled by locks and dams, the standard-size flotilla is 15 barges and one towboat. On the Lower Mississippi, where the river is free of locks, flotillas can reach a half mile (0.8 km) in length, with 40 or more barges offering a cargo capacity of over 58,000 tons (52,000 metric tons).

As the flotilla grows in size, so does the amount of horsepower required from the towboat. The standard for a towboat operating on the Upper Mississippi is 6000 hp (4.5 MW). Heavy-weight-class towboats operating on the Lower Mississippi, where the load is much larger, have up to 10,500 hp (8 MW) and are about 200 ft (60 m) long.

The design of a towboat is vastly different from that of a tug. The towboat (**Fig. 4**) has a relatively flat bottom with a square bow on which heavy, upright towing knees are fixed. Towboats have conventional doors and windows, and on some the upper decks are retractable for easy passage under bridges. Tugboats, on the other hand, have curved hulls and bows, and are designed with watertight doors and portholes to prevent the entry of water (**Fig. 5**).

Fig. 2. Roll on/roll off (RO/RO) deck barge. (*The American Waterways Operators*)

Fig. 3. Inland tow, consisting of open and closed dry cargo barges, being pushed ahead by a towboat. (*The American Waterways Operators*)

Fig. 4. Inland towboat. (*The American Waterways Operators*)

Fig. 5. Articulated tug-barge used in coastal towing. (*The American Waterways Operators*)

Barges and towboats cannot withstand the wind and wave action encountered on open water. Therefore, on open channels, such as the Atlantic and Gulf intracoastal waterways, tugboats are used, which pull barges on a long towline, or hawser. The number of barges that can be handled by this method is less than can be handled in push-towing operations; however, barges involved in coastal towing often greatly exceed the inland barges in size. The state-of-the-art in coastal towing is the articulated tug-barge, intended to increase efficiency and safety by eliminating the long towline. Instead, the tugboat fits into a notch in the stern of the barge with the two units tightly connected. This allows for more control in steering the barge.

Carrying capacity. One barge carries up to five times its weight in freight. One 1500-ton (1350-metric-ton) barge load of soybeans equals 1167 acres (472 hectares) of farm production at 45 bushels per acre (4 m^3 per hectare). In relation to other modes of transportation, one 24-barge flotilla has the cargo capacity of 180 rail cars or 1440 trucks.

Cargo. Many of the cargoes carried by barge are essential materials, such as petroleum, coal, iron and steel, chemicals, forest products, fertilizer, sand, and salt. Most commodities shipped by barge are those that lend themselves to bulk shipments and efficient cargo-handling techniques. Petroleum and petroleum products account for the largest percentage (31.9%) of cargo shipped by barge in the United States. Coal is second-largest, representing 25% of barge tonnage, followed by crude materials (forest products, minerals, ores, sand, stone, and gravel) and then food and farm products. Barges also transport ethanol, a new commodity market created by demand for alternative fuels. Other products shipped by barge include industrial chemicals, iron and steel products, sulfur, fertilizer, logs, and paper products. Cumbersome cargoes that cannot easily be shipped by other means are often moved by barge, including nuclear reactors and drilling platforms.

Barges transport over 60% of the nation's export grain, a significant amount of the coal that is delivered to domestic power plants and utilities (enough to produce 10% of the electricity used each year in the United States), and most of New England's home heating oil and gasoline.

Barge transportation is an important link in the nation's intermodal transportation network. For example, petroleum products are transported from Gulf Coast refineries by pipeline to New York and barged to New England power plants, where they are used to provide electricity, heat homes, and keep industry moving. Dry whey is trucked to Lewiston, Idaho, barged to Portland, Oregon, and shipped to Asia. Food, equipment, and supplies from the lower 48 states are moved by rail to Seattle, barged to Alaska, and trucked on the Alaskan Highway to the oil fields of the North Slope. The strength of these kinds of intermodal links to transport cargo is critical to the flow of the nation's commerce.

Waterways. Except for the Lower Mississippi, the Missouri, the Gulf Coast, and the Atlantic and Pacific waterways, the navigable rivers are slack water routes which have been improved for navigation by dredging and the construction of locks (**Fig. 6**) and dams. The St. Lawrence Seaway is a deep draft waterway that connects the Atlantic Ocean to the Great Lakes. It contains 13 Canadian and two United States locks. With the exception of the New York State Barge Canal, which is state-run, the inland waterways of the United States are maintained on a federal level by the Army Corps of Engineers. This includes the maintenance and modernization of the inland waterways locks and dams, the infrastructure that enables waterborne commerce. This infrastructure is critical, as failure of any of the aging locks on the Ohio River System or the Upper Mississippi River–Illinois Waterway System, for example, affects the flow of commerce on the entire interconnected waterways system. Many of these locks were built over 50 years ago when even the most optimistic projections did not match the traffic volume of today. Back-ups at these antiquated structures are common, and attention is being turned toward replacement and repair of these facilities. Since 1980, barge operators have paid a fuel tax that finances a percentage of the construction, operation, and maintenance costs for inland waterway projects. Before this time, all projects were entirely federally funded with appropriations from Congress.

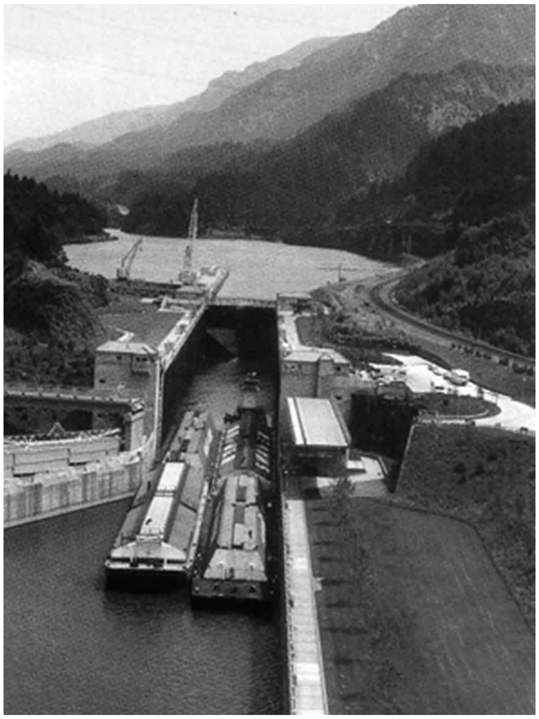


Fig. 6. Barge tow going through a lock in the Pacific Northwest. (*The American Waterways Operators*)

Most of the navigable inland rivers are open to navigation year-round, with the exception of the northern waterways. Commonly, ice conditions close the Upper Mississippi, portions of the Ohio River System, the Missouri River, and the New York State Barge Canal from December through March. Occasionally, ice impedes navigation on the other waterways as well. *See* CANAL; DAM; RIVER ENGINEERING.

Thomas J. Allegretti; Joseph Farrell

Bibliography. J. Leventhal, *Tugs: The World's Hardest Working Boats*, Black Dog & Leventhal Publishers, New York, 1999; *National Transportation Statistics*, U.S. Government Printing Office, annually; W. Salinger, *Tugboats 'n Towlines: The Men and Women Who Give Them Life*, Twin Lights Publishers, Rockport, MA, 2003; U.S. Army Corps of Engineers, *Waterborne Commerce of the United States*, 2004.

Inoculation

The process of introducing a microorganism or suspension of microorganism into a culture medium. The medium may be (1) a solution of nutrients required by the organism or a solution of nutrients plus agar; (2) a cell suspension (tissue culture); (3) embryonated egg culture; or (4) animals, for example, rat, mouse, guinea pig, hamster, monkey, bird, or human being. When animals are used, the purpose usually is the activation of the immunological defenses against the organism. This is a form of vaccination, and quite often the two terms are used interchangeably. Both constitute a means of producing an artificial but active immunity against specific organisms, although the length of time given by such protection may vary widely with different organisms. *See* VACCINATION.

Inoculation is the natural process of acquiring protection against disease, in that most persons are exposed to some organisms at times when no severe symptoms are displayed. The protective mechanisms of the body, especially antibody production, are stimulated by such a mild or insignificant exposure. An example of this is the discovery that the majority of adults have antibodies to poliomyelitis present, despite the absence of a history of the severe or recognizable disease form.

Inoculation may also refer to the deliberate seeding of organisms into culture media, and the introduction of fermenting bacteria, yeasts, or molds into various industrial processes that employ the chemical reactivity of these organisms. *See* BACTERIOLOGY; CULTURE; IMMUNITY; INDUSTRIAL MICROBIOLOGY.

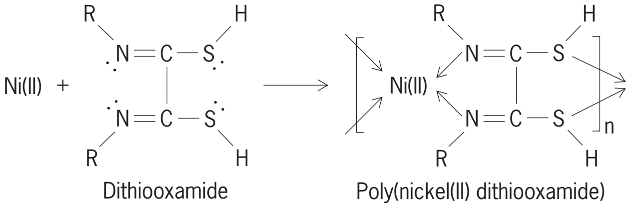
Edward G. Stuart; N. Karle Mottet

Inorganic and organometallic polymers

Macromolecules that contain metals, metalloids, or other elements. The inorganic element may lie along the polymer backbone where it is covalently bonded to the adjacent atoms (**Fig. 1***a, e*); the inorganic element may be ligand-coordinated (Fig. 1*b, c*); or the inorganic element may be pendant to the polymeric chain (Fig. 1*d*). Organometallic polymers are a special class of inorganic polymers in which the metal, metalloid, or other element is covalently bonded to carbon (Fig. 1*d–f*).

Many examples of inorganic polymers are found in nature having important physiological functions such as respiration, photosynthesis, energy transfer, and metal-ion storage. Small proteins called metallothioniens store metals such as copper and zinc in higher animals. For example, cd$_7$-metallothionein-2 is a rat protein rich in cysteine that is bonded to four zinc atoms through sulfur-containing amino acids. Ferrodoxins are bacterial proteins containing up to eight iron-sulfur complexes that are responsible for biological oxidation-reduction and electron transfer reactions. In addition, mammalian ferritin (a high-molecular-weight protein containing approximately 4500 iron atoms) is essential for the storage and transfer of Fe(II) and oxidation to Fe(III).

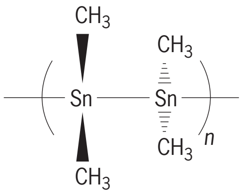
Synthetic inorganic and organometallic polymers were developed in the early part of the twentieth century. One of the first useful inorganic coordination polymers was prepared by impregnating a plastic film with dithiooxamide (a tetradentate ligand), and treating the film with a solution of Cu(II) or Ni(II) salts. The resulting linear polymer [reaction (1)] had light-polarizing properties. In the 1960s, new developments in synthetic methods led to a number of polymers that contain main-group, transition, and inner transition elements. Many of these materials were designed to withstand high temperatures and resist erosion, while maintaining desirable attributes such as oxidative stability, elasticity, electric conductivity or resistivity, optical properties, fire retardancy, catalytic properties, low density, high tensile strength, and inertness to chemicals.



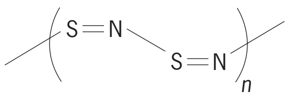
(1)

Classification. Inorganic and organometallic polymers may be classified in terms of morphology (chains, sheets, and networks) or backbone elemental composition (homoatom and heteroatom).

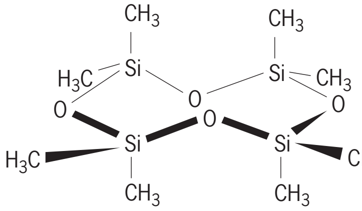
Chain polymers. A chain polymer is a one-dimensional (1D) macromolecule composed of linear repeating units that lack inter- or intrachain crosslinks. Examples are poly(dialkylstannane) (structure **1**) and polythiazyl (**2**). Under certain circumstances, the ends of linear chains connect to form macrocyclic structures. Thus, cyclic poly(methylhydrosiloxane) with as many as 50 —CH$_3$SiH— repeat units is known. Cyclic chains with a relatively small number of repeat units (oligomers) are very common. Some well-known cyclic oligomers are octamethylcyclotetrasiloxane, (Me$_2$SiO)$_4$ (**3**), and several allotropic forms of sulfur, for example, S$_8$ (**4**). Linear 1D chain polymers are also known to assume random coil, helical (proteinlike), and rigid rod structures.



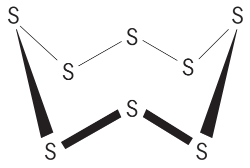
(1)



(2)



(3)



(4)

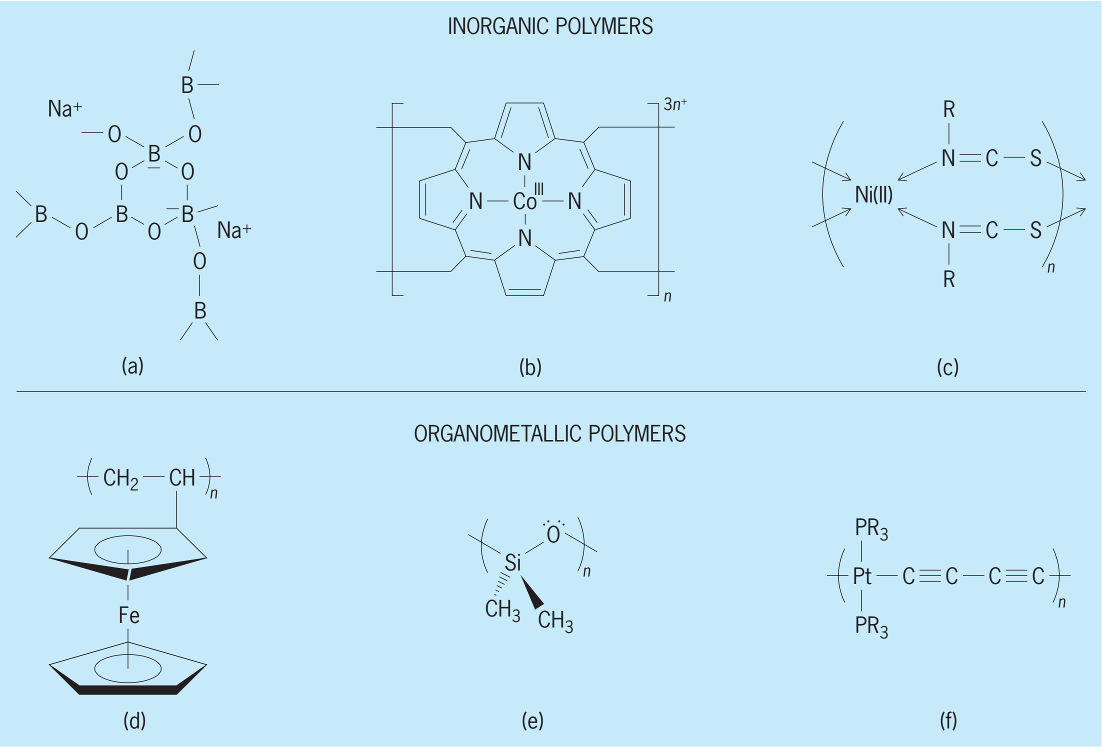


Fig. 1. Inorganic polymers: (*a*) borate polymer (metalloid in the backbone), (*b*) cobaltoporphyrin polymer (ligand-coordinated metal), and (*c*) nickel(II) dithiooxamide polymer (ligand-coordinated metal). Organometallic polymers: (*d*) ferrocene polymer (metalloid in backbone), (*e*) poly(dimethylsiloxane) (metalloid in backbone), and (*f*) platinum-diacetylene polymer (metal in backbone).

Sheet polymers. Sheet polymers are two-dimensional (2D) macromolecules in which the primary valences of the metal/metalloid atoms are satisfied by covalent bonds. Boron nitride (α-BN) is an example of a 2D sheet polymer (**Fig. 2***a*) in which each atom is bonded to three other atoms in the same layer. The stacked layers (lamellae) interact through van der Waals forces. Another variation of a 2D sheet occurs if two chains are crosslinked at regular intervals to give a ladder structure such as the crosslinked siloxane polymer (Fig. 2*b*). Ladder polymers are unusually high-melting solids with oxidative, hydrolytic, thermal, and chemical stability.

Network polymers. Network polymers are highly crosslinked materials in which the metal/semimetal atom valences are satisfied by bonds that result in a three-dimensional (3D) structure. Such polymers are usually difficult to characterize because of their refractory (ceramic-like) nature. Network polymers are often hard, infusible, and insoluble substances such as silicon dioxide and silicon nitride (Fig. 2*c*). These materials are added to organic polymers to enhance their strength and thermal stability.

Homoatom polymers. Poly(diphenylsilane), $(Ph_2Si)_n$, is an example of a homoatom polymer because it contains only silicon atoms along the backbone. Some

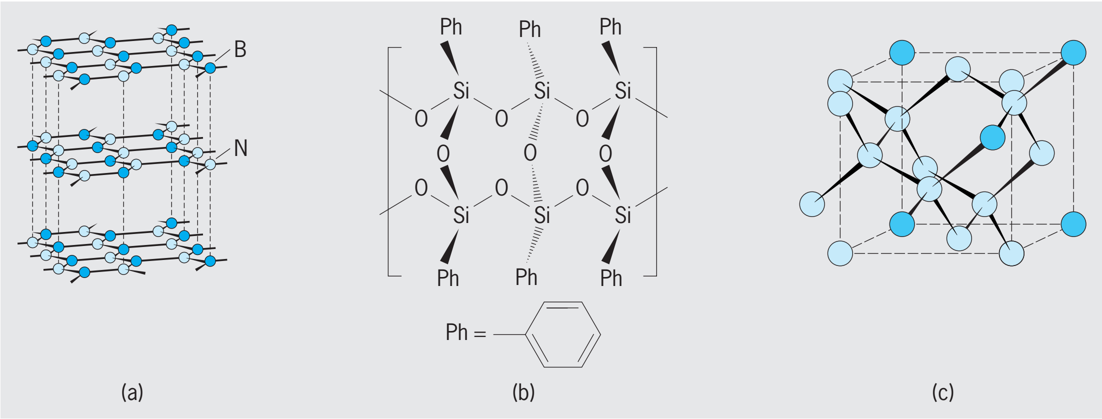
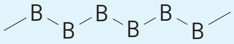
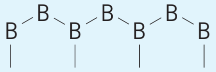
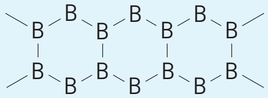
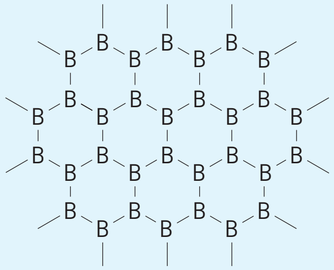
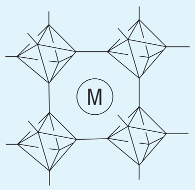


Fig. 2. Morphological classes of polymers: (*a*) boron nitride (α-BN), 2D parallel sheet; (*b*) poly(diphenylsiloxane), 2D ladder; and (*c*) silicon dioxide (silica), 3D network.

TABLE 1. Polymeric metal boride structures

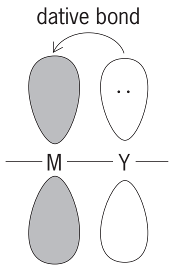
Formula	Metal	Morphology	Anion structure
MB	Fe, Ni	Zigzag chains	
$M_{11}B_8$	Ru	Branched chains	
M_3B_4	Cr, Ta	Ladder	
MB_2	Mg, Ti, Zr, Gd	Sheet	
MB_4	La	Network (octahedron)	 M = metal; B at each vertex of octahedra

metalloid anions have homoatomic polymeric structures. For instance, metal borides of formula M_xB_y contain the anion as chains, ladders, 2D sheets, or 3D networks (**Table 1**). These anionic polymers possess high thermal stability, excellent chemical resistance, and potentially useful electrical properties.

Certain elements, such as phosphorus, prefer to aggregate as low-molecular-weight oligomers. For example, white phosphorus is a tetramer (P_4) with phosphorus atoms at the corners of a tetrahedron. Black phosphorus, which is obtained by heating P_4 under pressure [200°C (390°F), 12,000 atm (1220 MPa)] with a catalyst, is a sheet polymer with each phosphorus atom bound to three neighbors in double layers. This giant molecule consists of stacked double layers, analogous to the boron nitride lamellae. In addition, black phosphorus is a good electrical conductor. Red phosphorus is made by heating P_4 at 400°C (750°F) in a sealed vessel. One crystalline form of red phosphorus has a linear, tubular arrangement of eight phosphorus atoms in a pentagonal wedge structure.

Elemental sulfur is also a polyatomic and polymorphic substance. Oligomeric sulfur rings consisting of 6, 7, 8, and 12 sulfur atoms have been identified. The orthorhombic α-S_8 form is the most thermodynamically stable and consists of eight-member puckered rings in a crown conformation stacked together in "crankcase" fashion. The chain polymer form of sulfur (catena or plastic sulfur) is obtained by quenching molten S_8 in water. If stretched, catena sulfur has a helical structure. Selenium forms similar polymers.

Heteroatom polymers. Heteroatom polymers contain a metal or semimetal in addition to one or more other atoms (such as oxygen, sulfur, nitrogen, phosphorus, and carbon) in the backbone. In general, polymer stability and inertness increases when the heteroatom is electron-rich. Presumably, such atoms serve as electron sources toward low-lying vacant orbitals on the metal or metalloid (structure **5**). The resulting



M = metal, semimetal
Y = heteroatom with a lone pair of electrons

(5)

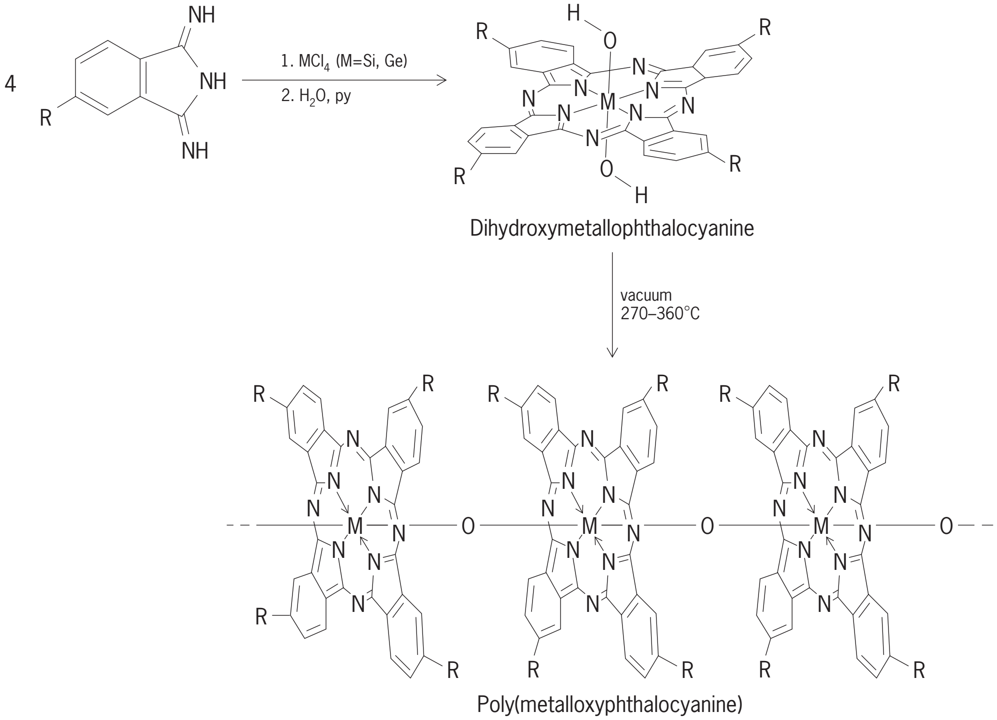
back-bonding (dative bond) generally decreases the electrophilicity of the metal or metalloid and reduces the susceptibility of the polymer to attack by Lewis bases. In addition, the interaction tends to stabilize the polymer toward thermal degradation and aids in the regulation of structure, chain mobility, and interchain forces that enhance its dynamic and mechanical properties.

Synthetic methods. A number of methods have been developed to synthesize inorganic and organometallic polymers. Representative examples of the more important and commonly used preparative techniques include step-growth (condensation) polymerization, ring-opening polymerization, chain-growth (addition) polymerization, and reductive (Wurtz-type) coupling polymerization.

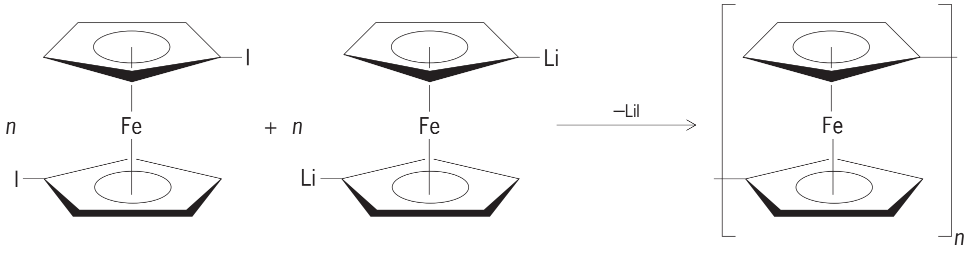
Step-growth polymerization. In step-growth polymerization, two molecules combine with the evolution of a small molecule such as water, ammonia, or a salt (for example, hydrogen chloride). For example, dihydroxymetallophthalocyanine [prepared from a metalloid halide such as silicon tetrachloride ($SiCl_4$) or germanium tetrachloride ($GeCl_4$) and phthalocyanine followed by hydrolysis] undergoes dehydration under vacuum at elevated temperatures to give a linear, 1D "shish-kebab" polymer with an —M—O—M—O— chain backbone [reaction (2)]. This material has excellent thermal and chemical stability, solubility in strong acids, interesting magnetic and optical properties, as well as electrical (semiconducting) properties when doped with iodine.

Another step-growth polymerization uses an inorganic or organometallic compound that contains a difunctional organic ligand. For example, diiodoferrocene reacts with a dilithioferrocene [reaction (3)] to give a polyferrocene. The driving force for the reaction is the precipitation of the salt, lithium iodide. Similarly, condensation with bridging difunctional reagents, such as the sodium salt of an organic diol, can produce metal-containing polymers [reaction (4)].

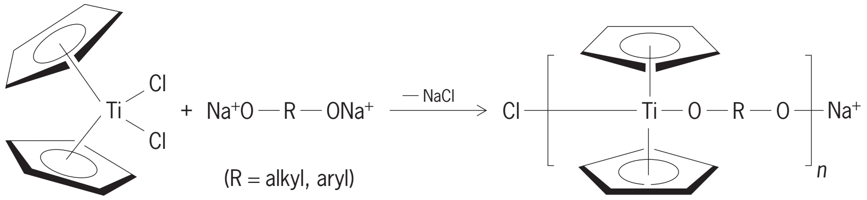
Ring-opening polymerization. Ring-opening polymerization is a principal method of preparing inorganic and organometallic polymers. The method is convenient and usually requires heat, light, or a catalyst. The reaction yields high-molecular-weight polymers without condensation products. An example of a ring-opening polymerization is the synthesis of polysiloxanes (silicones), which have many



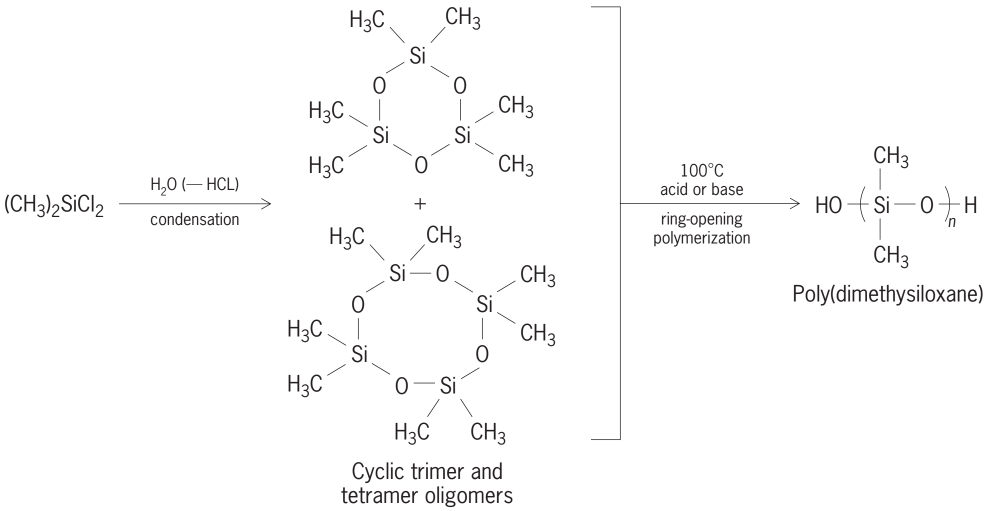
(2)



(3)



(4)



(5)

industrial and consumer applications, including:

Insulators
Dielectric materials
Heat-exchange fluids
Seals and gaskets
Caulking agents
Emulsifying agents
Surgery implants
Mold-forming agents
Antifoaming agents

Masonry additives
Surfactants
Paper release coatings
Coupling agents
Water repellants
Lubricants
Hydraulic fluids
Ceramic composites
Pressure-sensitive adhesives

The industrial process involves the hydrolysis of dimethyldichlorosilane followed by acid- or base-catalyzed ring-opening polymerization [reaction (5)] to give poly(dimethylsiloxane). The ring-opening polymerization method permits excellent control of molecular weight and polydispersity (molecular-weight distribution). **Table 2** summarizes other examples of rings and polymers that can be prepared by ring-opening polymerization.

Chain-growth polymerization. Chain-growth polymerization is initiated with free radicals, which may be generated chemically, thermally, photochemically, electrochemically, as well as with ions (cations or anions) or coordination complexes (for example, Ziegler-Natta or ferrocene catalysts). Invariably, chain-growth polymerization requires that the monomer contain an unsaturated functionality, usually an olefin. As a result, the polymers contain carbon in the chain backbone. Some examples of chain-growth polymerization that incorporate metals or metalloids into the polymer are given in **Fig. 3**.

Reductive (Wurtz-type) coupling polymerization. Reductive coupling polymerization uses active metals (such as sodium or potassium) in a reaction with difunctional, halogen-containing organometallic compounds (for example, organosilanes, organogermanes, organostannanes, carbometallanes, and ferrocenes). Typically, the reactions are carried out at a temperature that melts the alkaline metal and disperses it in a nonreactive solvent such as toluene or tetrahydrofuran. Alternatively, the reducing agent is made soluble with a chelating (complex-forming) agent such as crown ether.

For background information *see* BORON; COORDINATION COMPLEXES; GERMANIUM; INORGANIC POLYMER; LIGAND; METALLOID; NONMETAL;

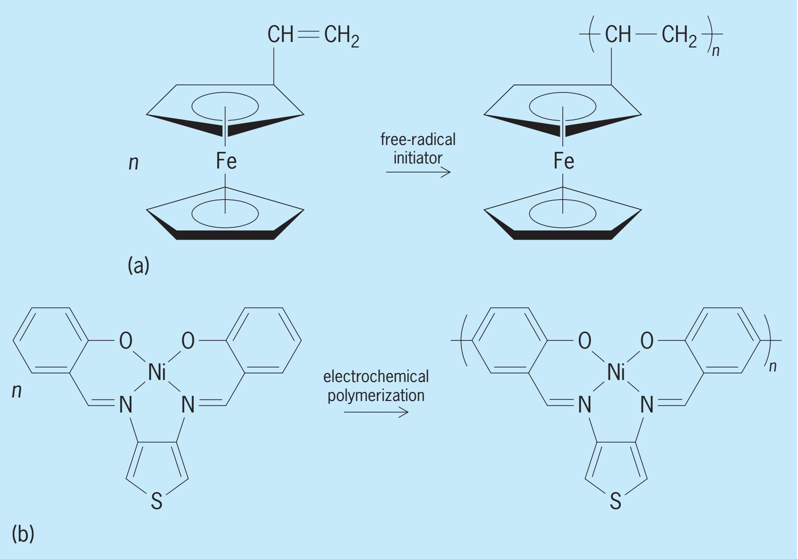
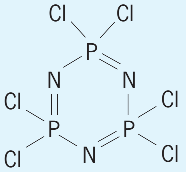
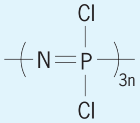
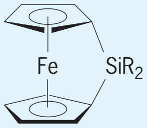
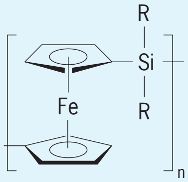
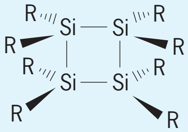
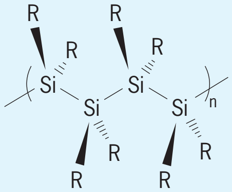
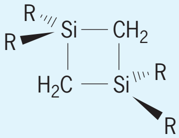
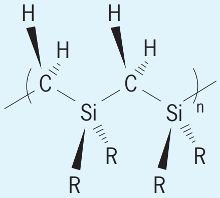


Fig. 3. Chain-growth polymerization synthetic routes. (*a*) Free-radical initiated. (*b*) Electrochemical polymerization.

TABLE 2. Some rings and polymers formed by ring-opening polymerization

Ring	Polymer	Name
		Polyphosphazene
		Poly(diorganosilyl ferrocene)
		Poly(diorganosilane)
		Poly(diorganocarbosilane)

ORGANOMETALLIC COMPOUND; POLYMER; POLYMERIZATION; SILICON; SULFUR in the McGraw-Hill Encyclopedia of Science & Technology. Martel Zeldin

Bibliography. R. D. Archer, *Inorganic and Organometallic Polymers*, 2001; I. Bertini et al. (eds.), *Bioinorganic Chemistry*, 1994; V. Chandrasekhar, *Inorganic and Organometallic Polymers*, 2005; J. E. Mark, H. R. Allcock, and R. West, *Inorganic Polymers*, 2d ed., 2005; C. U. Pittman, Jr., et al. (eds.), *Metal-Containing Polymeric Materials*, 1996.

Inorganic chemistry

The chemical reactions and properties of all the elements in the periodic table and their compounds, with the exception of the element carbon. The chemistry of carbon and its compounds falls in the domain of organic chemistry. The boundaries of inorganic chemistry with the other major areas of chemistry are not precisely defined, and it is often a matter of taste as to whether a particular topic is to be included in the field of inorganic chemistry or is to be considered physical or even organic chemistry. Physical chemistry may be defined as the application of quantitative and theoretical methods to chemical problems, and is a methodology rather than a specific body of knowledge. Investigations into theoretical inorganic chemistry or the study of problems in inorganic chemistry by quantitative and sophisticated physical methods may be considered either inorganic or physical chemistry quite arbitrarily. In similar fashion, organometallic compounds may be considered to be in the sphere of either inorganic or organic chemistry. The modern inorganic chemist is concerned with problems that once were considered the prerogative of physical chemists, organic chemists, or even biochemists. *See* PHYSICAL CHEMISTRY.

Synthetic inorganic chemistry. The reactivity of the elements of the periodic table varies enormously, and over a much wider range than is encountered in organic chemistry. Consequently, the inorganic chemist must frequently employ unusual apparatus and techniques. The elements range from the rare gases, which are unreactive and form very few chemical compounds, to the extremely reactive halogens and alkali metals. Fluorine is perhaps the most reactive element known; it forms compounds with all other elements, including the rare gases. Because of the great reactivity of fluorine and the closely related halogen fluorides, special metal and plastic apparatus must be used in experimentation. Both fluorine and the important compound hydrogen fluoride attack glass, and this common material of construction cannot be used for experimentation with these substances. There are methods to effect direct reaction between elemental fluorine and organic compounds without completely destroying the organic reactant, and this procedure has greatly increased the number of fully fluorinated organic compounds that can be readily obtained. *See* FLUORINE; FLUOROCARBON; HYDROGEN FLUORIDE; PERIODIC TABLE.

Another element important in synthetic inorganic chemistry since World War II is boron. The hydrides of boron were first obtained by reaction of a metal boride with a solution of an aqueous acid. The procedure was very difficult and tedious, and it required weeks or months of labor to obtain a few cubic centimeters of the gaseous product, which turned out to be the simplest boron hydride, B_2H_6 (diborane). Many years later it was discovered that yields could be increased by passing boron trichloride and hydrogen through an electric discharge, but the yields were still distressingly low. Under the impetus of wartime urgency, chemical syntheses for the boron hydrides were developed. Using the readily available alkali metal hydrides and boron trifluoride as starting materials, boron hydrides were obtained in very good yield. The metal borohydrides, such as $Al(BH_4)_3$ (bp 44.5°C or 112.1°F), when they can be prepared, are likely to be the most volatile compounds of the metal. *See* BORANE; BORON.

The carboranes are a relatively new class of organoboron compounds with a rich chemistry and many potential applications. Hydroboration, the addition of diborane to organic compounds containing a carbon-carbon double bond, is a preparative reaction that has many important applications in organic synthesis. Many compounds of elements such as silicon, germanium, gallium, phosphorus, and nitrogen, some of which were at one time considered too unstable to be synthesized, have been prepared by the synthetic inorganic chemist. An outstanding achievement in modern inorganic synthesis was the preparation of the noble gas fluorides. The discovery that the noble gases could form compounds with fluorine and oxygen resulted in a whole new area of inorganic research. *See* CARBORANE; HYDROBORATION; INERT GASES.

Many boron compounds react violently with air and water. It is necessary, therefore, to use all-glass vacuum systems for carrying out experiments with these substances. Vacuum-line techniques are widely used in inorganic chemistry for the manipulation of volatile, highly reactive compounds such as the hybrides of phosphorus, silicon, and related compounds, and an array of vacuum lines often is the hallmark of a laboratory for synthetic inorganic chemistry.

The synthesis of inorganic compounds as an end in itself, which at one time was perhaps the most characteristic of the activities of the inorganic chemist, has been largely superseded by synthesis directed to specific goals; among these are compounds required to elucidate fundamental laws of chemical bonding, reactivity, and structure; inorganic polymers with unusual thermal and mechanical properties; compounds or systems with superconductor or electron conductor properties; and compounds with particular virtues as reagents in organic syntheses.

Coordination chemistry. The transition elements in the periodic table form coordination compounds with a great variety of organic donor molecules. Traditionally, coordination chemistry has been a major interest of inorganic chemists. However, interest in maintaining a large flow of new coordination compounds whose principal virtue is that they can be easily prepared has considerably diminished in most quarters. The principal interest in coordination chemistry is in characterizing the physical properties of coordination compounds with unusual properties or uses. Thus, the synthesis of multidentate ligating agents such as the crown ethers and the cryptates, which have high specificity for binding particular ions, became a very important activity. Macrocyclic polyethers and macrobicyclic polyethers with tertiary amino groups at the bridgeheads mimic many of the properties of ion-transporting agents in living systems. The template properties of transition-metal ions for the synthesis of multidentate ligands such as the porphyrins and corrins that occur in nature have also received considerable attention. These and other initiatives in coordination chemistry have become an important component in the discipline of bioinorganic chemistry. *See* COORDINATION CHEMISTRY; MACROCYCLIC COMPOUND.

Organometallic compounds. Organometallic compounds constitute a borderline area of study between inorganic and organic chemistry. Until 1950, the organometallic compounds known were entirely compounds of the principal families in the periodic table; no stable derivatives of subgroup metals or transition metals had been prepared. Today a very large number of organometallic compounds of the transition and subgroup elements with metal-carbon bonds to cyclopentadiene, C_5H_6, aromatic hydrocarbons such as benzene, C_6H_6, as well as other organic compounds, have been synthesized. These compounds form interesting derivatives with carbon monoxide, CO, and can also undergo numerous reactions, and the structure and properties of these compounds pose interesting theoretical problems. Homogeneous catalysis by organometallic compounds has become one of the most active areas of catalysis research. Hydride complexes of ruthenium, rhodium, and iridium are catalysts for hydrogenation, olefin isomerization, hydroformylation, and olefin polymerizations. Much effort has been devoted to the development of homogeneous organometallic catalysts for the reduction of carbon monoxide by hydrogen; these catalysts are intended to be more selective and to function under milder conditions than those required in heterogeneous catalysis. Organometallic reagents are also being sought for highly selective syntheses of complicated molecules such as pharmaceuticals employed in therapy. Palladium complex compounds are being used in the synthesis of alkaloids, and the preparation of such important classes of compounds as the insect pheromones and the prostaglandins is being facilitated by the use of nickel and aluminum organometallic compounds. *See* ALKALOID; EICOSANOIDS; HOMOGENEOUS CATALYSIS; ORGANOMETALLIC COMPOUND; PHEROMONE.

Solid-state chemistry. The elements and compounds that form the subject matter of inorganic chemistry exhibit a very wide range of physical properties which run the gamut from helium, the substance of lowest known melting and boiling point, to elements such as tungsten, titanium, and carbon, which are among the most refractory high-melting substances known. The study of solid-state reactions and compounds that can be prepared by them has assumed great importance, and is another major research area in inorganic chemistry. Unlike gaseous compounds and most organic compounds, which obey the law of definite proportions, many solid compounds, particularly those of the transition elements, exhibit variability of composition, or as it is frequently designated, nonstoichiometry. When a solid compound deviates from simple stoichiometric relations, it contains an excess of either positively charged metal cations or negatively charged anions. Such solid systems frequently show unusual electronic properties, which are made use of in many solid-state devices. Photovoltaic cells, for example, which convert light to electricity, belong to this group of nonstoichiometric solid-state devices. *See* NONSTOICHIOMETRIC COMPOUNDS; REFRACTORY.

Transistors, thermistors, phosphors, and light-emitting diodes are also important solid-state devices. Because these materials are frequently prepared by solid-phase reactions, high temperatures may be required for the reaction. High-temperature chemistry has greatly expanded in scope, not only in the preparation of nonstoichiometric compounds, but also in the preparation of refractories useful at very high temperatures in space and nuclear technology. Although many reactions can be forced to proceed in the desired direction by only increasing the temperature, this is not always adequate in all instances. Another dimension in solid-state chemistry was added by the simultaneous use of high temperature and extremely high pressures. With the equipment available, it is possible to carry out solid-state reactions at temperatures of approximately 2500°C (4532°F) and pressures of the order 10^5 millipascals (10^6 bars). Under conditions of high temperature and pressure, ordinary carbon (graphite) can be converted into diamond, and since the pressure is actually sufficient to distort the electron orbitals, new varieties of matter can be prepared. For example, liquid hydrogen has been converted to a metallic form at these ultrahigh pressures. *See* HIGH-PRESSURE CHEMISTRY; HIGH-TEMPERATURE CHEMISTRY; LIGHT-EMITTING DIODE; THERMISTOR; TRANSISTOR.

The solid-state systems used for semiconductors, photovoltaic devices, and laser photodiodes have traditionally been highly ordered, crystalline materials. However, substances in the amorphous or glassy state have been found to possess very interesting properties for similar uses. Amorphous silicon appears to have considerable potential in solar energy conversion cells for direct production of electricity, and amorphous silicon and germanium have semiconductor properties quite analogous to their crystalline counterparts. Amorphous vanadium or molybdenum sulfides can store lithium ions in nonaqueous lithium storage batteries better than crystalline materials can. Amorphous systems are only imperfectly understood, but they possess much scientific and technical interest and can, therefore, be expected to attract continued scientific investigation. *See* AMORPHOUS SOLID; METALLIC GLASSES; SEMICONDUCTOR.

Layered materials were discovered that have unusual properties, and this discovery revived interest in the synthesis and study of layered compounds. Graphite intercalation compounds, for example, consist of layers of carbon atoms that have halogen or other atoms or molecules inserted between the carbon layers. Certain of these intercalation compounds have an electrical conductivity comparable to that of metallic copper, and the possibility that such systems could replace metallic electrical conductors has aroused much interest. Intercalated room-temperature superconductors are also being sought. Intercalation compounds derived from graphite or from transition metal sulfides are effective and highly selective reagents in organic synthesis. Many layered compounds can accommodate layers of ions with high mobility. These systems are in effect solid electrolytes or charge carriers and may contribute significantly to the long-sought lightweight high-capacity electric storage battery.

Still another class of electronic conductors comprises chains and clusters of metal atoms. Linear chains of transition-metal atoms (Krogmann salts) are one-dimensional electron conductors. Metal cluster compounds also have extremely interesting electrical properties. Organometallic compounds containing metal clusters with as many as 28 platinum metal atoms and 44 carbonyl ligands have been made. The metal atoms in the cluster are directly bonded to each other and exhibit many of the properties of solid metal catalysts. The clusters often exhibit superconductivity at higher than customary temperatures, and have, therefore, become a point of departure in the search for new superconductive materials that may be suitable for electric power transmission at liquid hydrogen rather than liquid helium temperatures. Metal cluster chemistry ranks among the most promising of the approaches being explored in the search for new catalysts. *See* METAL CLUSTER COMPOUND; SOLID-STATE CHEMISTRY; SUPERCONDUCTIVITY.

Geochemical aspects. Many of the synthetic procedures in inorganic chemistry carried out at high temperatures and pressures have considerable interest in geochemistry. Some of the earliest inorganic chemistry was practiced in connection with mineralogy. Mineral syntheses, or the preparation of inorganic compounds identical with those found in nature, provide important information to the geochemist. Not only has such research helped to explain the sequence of chemical reactions and conditions responsible for the formation of minerals in nature, but many minerals and gems such as diamond, ruby, sapphire, quartz, and corundum are now manufactured on the large industrial scale. High-temperature reactions at very high pressures or hydrothermal reactions at high temperatures and pressure are employed for this purpose. With the advent of ultrahigh-pressure equipment, it became feasible to study chemical reactions under conditions approximating those many miles below the Earth's surface, and such studies are expected to add greatly to the understanding of geochemical phenomena. *See* GEM; GEOCHEMISTRY; HIGH-PRESSURE MINERAL SYNTHESIS; HIGH-PRESSURE PROCESSES.

Nuclear science and energy. The development of nuclear energy since World War II has provided a great impetus to inorganic chemistry. The discovery of the transuranium elements was one of the outstanding events in chemical science, and opened up an entirely new area of the periodic table for investigation. The chemistry of the actinide 5*f* elements has revealed many surprises, and despite the great amount of new chemistry that has already been acquired, much still remains to be learned about the very complex phenomena associated with the oxidation-reduction behavior, the ions in solution, and the metallic states of the actinide elements. A particularly active field of actinide element research has been the synthesis of organometallic compounds of the 5*f* elements. The solvent-extraction and ion-exchange procedures developed for separating the actinide elements have had widespread applications in other areas of inorganic chemistry. Nuclear technology has also provided the impetus for the development of other separation procedures, for example, the separation of zirconium and hafnium, and of the rare-earth elements from each other, and has generally served to reinforce the traditional interest of the inorganic chemist in separations procedures. The development of safe, effective processing methods for the intensely radioactive spent nuclear fuels from breeder reactors, and the solution to the problem of safe storage and disposal of nuclear waste, remain paramount challenges to the inorganic chemist. *See* ACTINIDE ELEMENTS; ION EXCHANGE; NUCLEAR CHEMISTRY; NUCLEAR FUELS REPROCESSING; RARE-EARTH ELEMENTS; SOLVENT EXTRACTION; TRANSURANIUM ELEMENTS.

Applications in organic chemistry. Many of the most important advances in organic chemistry since 1900 resulted from the introduction of inorganic substances as reagents. Synthetic reactions based on magnesium metal gave rise to the vast corpus of Grignard chemistry, and the investigation of the metal carbonyls provided the impetus for the development of acetylene chemistry. Hydroboration, a process for producing organoboranes, can be used for stereospecific syntheses, asymmetric syntheses, and isomerizations. Hydroboration must be reckoned as among the most versatile synthetic procedures added to the armory of the organic chemist. Other inorganic substances that have found important use in inorganic chemistry are selenium for dehydrogenation reactions; lead tetraacetate and thallium compounds for selective oxidations; aluminum chloride as a catalyst for alkylation, acylation, and ring-closure reactions; anhydrous hydrogen fluoride for diazotization, nitration, and sulfonation; and lithium aluminum hydride (and various of its derivatives) and alkali-metal borohydrides for selective reduction reactions. Metal or carbon atoms in the gas phase are remarkable reagents for gas-phase organic reactions. *See* GRIGNARD REACTION; METAL CARBONYL; ORGANIC CHEMISTRY; ORGANIC SYNTHESIS.

Reaction kinetics and mechanisms. Organic reactions generally proceed with the skeleton of the molecule remaining intact. Inorganic gas-phase reactions, on the contrary, are usually characterized by a complete disruption of molecular structure followed by reorganization to form the products of the reaction. Gas-phase reactions of inorganic compounds are thus in principle more complicated than are the usual organic reactions. Electron transfer in the oxidation-reduction reactions of transition-metal ions and compounds in solution has been a particularly important topic of theoretical interest. As is the case with electron transfer reactions in general, the role of electron tunneling in inorganic redox reactions still has many obscure features. This is particularly true of the very complex redox reactions of the actinide elements. Hydrolytic reactions of highly charged ions in aqueous solutions are also important in the mechanisms of many inorganic reactions and are important aspects of modern transition-element and actinide element chemistry. These subjects have become the concern of physical and theoretical chemists, but they also continue to be an important part of contemporary inorganic chemistry. *See* CHEMICAL DYNAMICS; OXIDATION-REDUCTION.

Bioinorganic chemistry. Inorganic chemistry is under pressure, as are other areas of chemical research, to escalate attention from simple molecules to more complicated systems. Like the organic chemist, the inorganic chemist has turned to biology as a new field to explore, and much of the most interesting contemporary work in organic chemistry is directed to the inorganic chemical aspects associated with living organisms.

Metal ions are very important participants in many biological phenomena. These range from the role of trace elements as essential nutrients to such questions as the function of metal ions in respiratory and photosynthetic pigments. Other important questions relate to the function of the metal ion in metal-containing enzymes involved in oxidation or reduction reactions, and to the transport of metal ions across membranes. Platinum and other metal coordination compounds have been investigated for possible use in the therapy of cancer. The basis for the use of lithium ion to control manic-depressive fluctuations in mood also poses important questions for the inorganic chemist.

A major activity in bioinorganic chemistry is the preparation and study of model systems intended to mimic the behavior of important biological entities. Thus, the construction of model systems for the prosthetic group of respiratory pigments (hemoglobin, myoglobin), oxidative enzymes (peroxidase), electron transfer proteins (cytochrome, ferredoxin, plastocyanin), and enzymes such as nitrogenase (the nitrogen-fixing enzyme) is under very active investigation by inorganic chemists.

The role of zinc ion in such enzymes as carbonic anhydrase and proteolytic peptidases has also focused attention on the inorganic components of these and other important enzyme systems. The application of modern spectroscopic techniques, particularly infrared and magnetic resonance spectroscopy, has suggested a role for the central magnesium atom of chlorophyll in the light-energy conversion step of photosynthesis. *See* BIOINORGANIC CHEMISTRY; SPECTROSCOPY.

Inorganic technology. The production of inorganic chemicals is a basic aspect of the chemical industry, and the heavy inorganic chemicals sulfuric acid, ammonia, chlorine, and phosphoric acid provide indispensable materials for many industries. The production of glass, ceramic, cement, fertilizer, and metals essentially involves reactions of inorganic chemistry carried out on the large scale. These industries have become less traditional in nature because of the contributions from inorganic chemical research, and it is altogether likely that modern inorganic chemistry will continue to make critically important contributions to the continued evolution of inorganic technology.

Joseph J. Katz

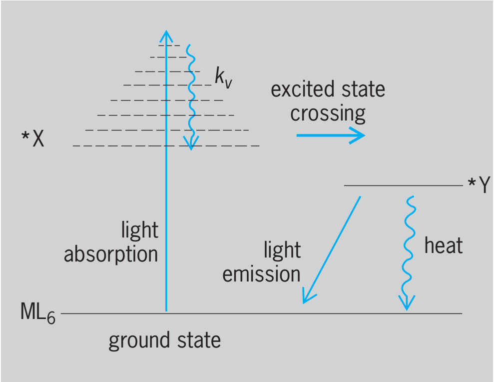
Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., 1999; G. L. Eichhorn and L. G. Marzelli (eds.), *Advances in Inorganic Biochemistry,* vol. 6, 1985; N. N. Greenwood and A. Earnshaw, *Chemistry of the Elements*, 2d ed., 1997; C. E. Housecroft and A. G. Sharpe, *Inorganic Chemistry*, 2d ed., 2004; J. E. Huheey, E. A. Keiter, and R. L. Keiter, *Inorganic Chemistry: Principles of Structure and Reactivity*, 4th ed., 1997; K. Mackay, A. Mackay, and B. Henderson, *Introduction to Modern Inorganic Chemistry*, 6th ed., 2002; G. L. Miessler and D. A. Tarr, *Inorganic Chemistry*, 3rd ed., 2003.

Inorganic photochemistry

The study of the light-induced behavior of various metal compounds. The physical and chemical properties of substances are generally altered by the absorption of light. Typical metal compounds have a characteristic number (coordination number) of molecules or ions (ligands) directly bonded to the metal center. Many of these compounds are colored, and it is possible that some metal compounds could mediate the transformation of solar radiation into useful chemical or electrical energy, or that they might scatter light in such a way that its frequency is increased. *See* LIGAND; NONLINEAR OPTICS.

The photochemistry of metal compounds has grown in concert with modern theories of the electronic structure of molecules and of chemical bonding in molecules; photochemical studies are often designed to probe and test these theories. The range of studies spans some of the subdisciplines of chemistry, and includes or bears on such topics as photophysics, the development of laser materials, catalysis, physics, photosynthesis, oxidation-reduction chemistry, acid-base chemistry, organometallic chemistry, metalloenzyme chemistry, solid-state chemistry, and surface chemistry. *See* CHEMICAL BONDING.

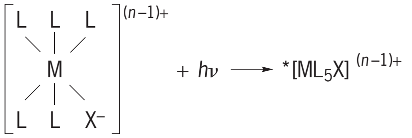
Excited states. The absorption of light results in a rearrangement of electrons within a molecule. In many molecules, the new electron configuration can persist for a significant period of time. Chemical properties depend on electronic configurations, and the ground-state and excited-state electronic configurations differ. In general, the bond lengths (angles and so on) of the excited state of a molecule will be different from those of the ground-state (or thermally equilibrated) molecule. Consequently it is useful to regard these excited states of molecules as new chemical species with chemical properties distinctly different from those of the ground state. Some of the energy used to generate an excited state is degraded to heat when the excited molecule relaxes to a bonding arrangement compatible with the new electronic configuration. In addition, the initial excited state (*X) may rapidly convert (or cross) to a lower-energy excited state (*Y) with yet another electronic configuration (see **illus.**). After light absorption in metal compounds, the time required for generation of the lowest-energy excited state is usually very short (less than a nanosecond). As a consequence, any chemistry due to the higher-energy excited states (*X; see illus.) must occur very quickly, and either is intramolecular or involves the nearest neighbors of the excited molecule in condensed phases. The lowest-energy excited states of some metal compounds (*Y; see illus.) can exist for nearly a millisecond. This is long enough for many



Qualitative electronic energy level scheme for a six-coordinate metal compound (ML_6). Absorption of light by ML_6 results in an excited state *(ML_6) in which electrons are rearranged (either *X or *Y). Any bond length difference between ML_6 and *(ML_6) results in the intial excited state being vibrationally excited (bonds stretched or compressed from the equilibrium position). Vibrational relaxation k_v is rapid in condensed phases.

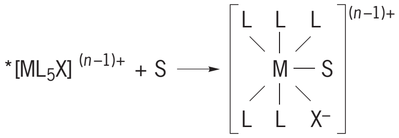
collisions in condensed phases, and such excited-state species can often react with other molecules present in the medium (bimolecular reactions). The lifetimes of excited molecules are limited by the probability of chemical reaction, by the probability of return to the ground state with the emission of light, and by the probability of return to the ground state without light emission but with the generation of heat energy. *See* ELECTRON CONFIGURATION; EXCITED STATE.

Unimolecular reactions. The initial steps of the simplest excited-state chemical reactions involve only the excited molecule or solvent molecules, and may be classified as excitation [reaction (1), where M is a metal and L and X are ligands]; decrease in coordination number [reaction (2)]; increase in coordination number for a solvent species [S; reaction (3)]; and homolysis, as in oxidation-reduction, in which bonding electrons shared equally in the products [reaction (4), with ·X as a free radical].

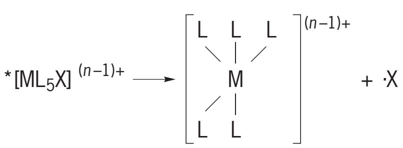


(1)

$$*[ML_5X]^{(n+1)+} \longrightarrow ML_5^{n+} + X^{-}(\text{or } L) \qquad (2)$$



(3)



(4)

Each of the chemical processes (2), (3), or (4) gives rise to an unstable, therefore reactive, chemical intermediate. Such intermediates are often useful in the synthesis of new compounds, as in reaction (5).

$$ML_5^{n+} + Y^{-} \longrightarrow [ML_5(Y^{-})]^{(n-1)+} \qquad (5)$$

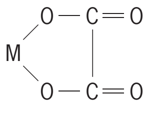
When these intermediates are generated on a surface or in an inert matrix, their lifetimes may be enhanced, and they may function as reactive catalytic sites. *See* CATALYSIS; FREE RADICAL; REACTIVE INTERMEDIATES.

The changes in the number of ligands coordinated to the central metal [reaction (2) or (3)] can be systematically treated by using theoretical models of the relationship between the electronic configuration of the molecular species and the nature of the coordinate covalent bond. The net chemical changes observed depend on the chemistry of the intermediate species $[(ML_5)^{n+}$ or $(ML_5SX)^{(n-1)+}]$ as well as the chemistry of the reactive excited states $[(^*ML_5X)^{(n-1)+}]$.

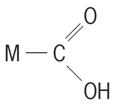
While excited-state reactions that result in changes in coordination number produce a single reactive species, the intermediate metallofragment, homolytic cleavage of a metal-ligand bond results in formation of a pair of very reactive species: a reduced metallofragment and a one-electron oxidized ligand species. These very reactive substances tend to recombine to regenerate the original metal compound, in addition to reacting with themselves or other solution species. The recombination reaction may be very rapid (in less than a nanosecond) when these species are formed in proximity, or it may be somewhat slower (on the order of 10^{-6} s) if the species manage to separate a few molecular diameters. In order for a simple homolytic process to be observed, the minimum energy difference between the reactive excited state and the ground state must exceed the M^{n+}-(X^{-}) bond energy. However, there are some systems in which the solvent can assist the homolytic process by means of a concerted displacement of the departing radical fragment. The net quantum yields for photohomolytic processes are reasonably large (that is, greater than 0.1) only when the recombination reactions are slow compared to other reactions of the reactive fragments. Homolytic reactions tend to dominate the chemistry of the excited states of cobalt(III) complexes.

Homolytic reactions turn out to be very efficient for methylcobalamin and other organocobalt complexes related to vitamin B_{12}. These materials are very highly colored (deep red), and most of the light photons absorbed result in homolytic cleavage of the cobalt-carbon bonds. Various recombination reactions decrease the observed product yields to less than 30%. This photosensitivity results in degradation of these natural complexes, and photoinduced reactions are not important to their enzymatic function. However, the small photonic energy required for cobalt-carbon homolysis is a manifestation of the weakness of this chemical bond. Thermal cleavage of the cobalt-carbon bond has been postulated as a key step in the enzymatic functioning of coenzyme-B_{12}.

Inevitably, many of the photoinduced unimolecular processes found for metal compounds cannot be neatly placed in the above categories. Among the most intriguing are processes that involve bond-breaking or bond-making processes on a ligand. For example, irradiations of some rhodium(III) or iridium(III) azide complexes (M^{III}-N_3^{-}) result in cleavage of the N^{-}-N_2 bond forming a coordinated nitrene, M^{III}-N^{-}. Another remarkable example is cleavage of the O_2C-(CO_2^{2-}) bond in some oxalate complexes (**1**) to form a metal-carbon bonded formate complex (**2**).

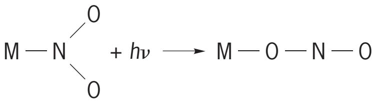


(1)



(2)

There are a number of photoinduced isomerizations [reaction (6)] that might appear to fall outside the categories listed above but in which the primary photoprocess appears to involve metal-ligand homolysis, followed by very rapid recombination of the fragments before they can be separated by one or more solvent molecules. *See* NITROGEN COMPLEXES.



$$\mathrm{M{-}NO_2} + h\nu \longrightarrow \mathrm{M{-}O{-}N{-}O} \qquad (6)$$

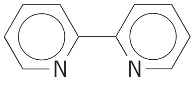
Bimolecular excited-state reactions. Reactions between excited states and molecules other than solvent species are very important when the electronically excited molecules are long-lived. Metal compounds are frequently facile oxidation-reduction reagents, and metal-compound excited states can be employed to displace oxidation-reduction equilibria. The important categories of bimolecular excited-state reactions are electron transfer [reaction (7*a*) or (7*b*)] and quenching by electronic energy transfer [reaction (8)].

$$^{*}[\mathrm{ML}_6^{n+}] + \mathrm{A} \longrightarrow \mathrm{ML}_6^{(n-1)+} + \mathrm{A}^+ \qquad (7a)$$

$$^{*}[\mathrm{ML}_6^{n+}] + \mathrm{B} \longrightarrow \mathrm{ML}_6^{(n+1)+} + \mathrm{B}^- \qquad (7b)$$

$$[\mathrm{ML}_6^{n+}]^{*} + \mathrm{C} \longrightarrow \mathrm{ML}_6^{n+} + \mathrm{C}^{*} \qquad (8)$$

The transfer of an electron from one species to another constitutes a net redistribution of electrical charge. As a consequence, there is necessarily some difference in the solvation of reactants and products; in addition, there will often be differences in the bond lengths of reactants and products. The larger these differences in bond length and solvation, the slower will be the rate of the electron-transfer process. Excited-state lifetimes are necessarily short, so that electron-transfer reactions must occur rapidly if they are to be of any consequence. Very rapid excited-state electron-transfer reactions occur most often between large molecules (thus minimizing solvation energies); most often between molecules in which the electron transfer does not result in large bond-length changes; and in part because the excitation energy stored in the excited molecules helps make the reactants appreciably less stable than the products. Typical of the metal compounds employed in the study of such reactions are *tris*-bipyridyl complexes, $\mathrm{M(LL)}_3^{n+}$, where LL equals structure (3) and



(3)

M^{n+} is Ru^{2+} or Cr^{3+}. Enough energy is released in the electron-transfer process to overcome some of the intrinsic factors limiting the rate. Indeed, typical metal-compound excited states can store enough energy that even the recombination reactions $\mathrm{ML}_6^{(n-1)+}$, $+ \mathrm{A}^+ \rightarrow \mathrm{ML}_6^{n+} + \mathrm{A}$ or $\mathrm{ML}_6^{(n+1)+} + \mathrm{B}^- \rightarrow \mathrm{ML}_6^{n+} + \mathrm{B}$, could ideally produce 1–2 V of electrical energy in a battery. Such excited-state electron transfer reactions afford a convenient means of generating very reactive intermediate species (A^+, B^-). *See* BOND ANGLE AND DISTANCE; ELECTRON-TRANSFER REACTION.

Quenching by electronic energy transfer is a theoretically complex process since it requires simultaneous relaxation of a donor electron and excitation of an electron in the acceptor molecule. Although such reactions tend to be degradative in most systems, they can be employed to generate chemically interesting acceptor-molecule excited states that cannot be easily populated by direct light absorption.

Photoinduced electron transfer and supramolecular clusters. Attempts to utilize metal compounds to mediate the transformation of light energy from the Sun into a useful chemical fuel have often focused on the cleavage of water, shown in reaction (9).

$$\mathrm{H_2O} + \mathrm{ML}_6^{n+} + h\nu \longrightarrow \mathrm{H_2} + {}^{1}\!/_{2}\mathrm{O_2} + \mathrm{ML}_6^{n+} \qquad (9)$$

Several metal-compound excited states store enough energy to promote this reaction as it is written. However, detailed consideration of reaction (9) reveals that it must involve at least two water molecules ($\mathrm{H_2O}$) and four electrons per product molecule formed; and the energy requirements for this reaction cannot be simply equated to the energy available from single-electron-transfer processes of molecular excited states [as in reactions (7*a*) and (7*b*)]. However, reactive intermediates generated in electron-transfer processes might be capable of transforming more than one equivalent of electrons (for example, by reduction of H^+ to a coordinated hydride H^-). Appropriate intermediates might be homogeneous (for example, low-valent complexes in solution) or heterogeneous (for example, colloidal metals or metal oxides).

Covalently linked aggregates of transition-metal complexes can be synthesized in which a specific photoreaction is favored. For example, one complex many function as a light absorber and sensitizer, and a second complex may function as the site of an excited-state reaction following intramolecular electron transfer. In principle, the acceptor center could be a semiconductor electrode or particle [for example, of titanium oxide ($\mathrm{TiO_2}$)] as well as a coordination complex. Some such clusters of light-absorbing complexes and acceptor centers can generate reasonably good photocurrents in solar cells, while others may be useful as efficient photocatalysts. In some such clusters, it is possible to accumulate sufficient charge separation to drive reactions as complex as (9). *See* CHEMICAL DYNAMICS; COORDINATION CHEMISTRY; COORDINATION COMPLEXES; LASER PHOTOCHEMISTRY; PHOTOCHEMISTRY; SUPRAMOLECULAR CHEMISTRY.

John F. Endicott

Bibliography. A. W. Adamson and P. D Fleischauer (eds.), *Concepts of Inorganic Photochemistry*, 1975, reprint 1985; V. Balzani and F. Scandola, *Supramolecular Photochemistry*, 1991; G. J. Ferraudi, *Elements of Inorganic Photochemistry*, 1988; M. A. Fox and M. Chanon

(eds.), *Photoinduced Electron Transfer*, 1988; V. Ramamurthy and K. S. Schanze (eds.), *Organic and Inorganic Photochemistry: Molecular and Supramolecular Photochemistry*, vol. 2, 1998; C. E. Wayne and R. P. Wayne, *Photochemistry*, 1996.

Inositol

The common name for 1,2,3,4,5,6-hexahydroxycyclohexane, which is classified as a carbohydrate. Inositol has nine different naturally occurring structural isomers. However, the most common form used biologically is D-*myo*-inositol (**Fig. 1***a*), commonly referred to as *myo*-inositol. Biologically, *myo*-inositol is used in the formation of the phosphatidylinositols in organisms, including plants, animals, and microorganisms. Phosphatidylinositols are used in a variety of cellular functions such as cell signaling, anchoring of proteins to the membrane, membrane trafficking, cellular development, and interaction with cytoskeletal elements. *See* CARBOHYDRATE.

Synthesis. The cellular biosynthesis of *myo*-inositol occurs by the cyclization of D-glucose-6-phosphate, an intermediate of carbohydrate metabolism, followed by dephosphorylation by a specific phosphatase. The cyclization reaction requires nicotinamide adenosine dinucleotide (NAD) as a cofactor and is catalyzed by the enzyme glucose-6-P cyclase. Mammals have a very limited ability to synthesize inositol, so it is considered an essential nutrient in their diets. *myo*-Inositol is commercially produced from phytic acid (*myo*-inositol hexaphosphate), which is found in corn-steep liquor following wet milling of corn. *See* CARBOHYDRATE METABOLISM.

Role in phosphatidylinositol synthesis. The most important biological function of *myo*-inositol is its incorporation into phosphatidylinositol (PI; Fig. 1*b*), a key phospholipid regulating cellular biochemical pathways. The biosynthesis of PI is a two-step process. In the first step, phosphatidic acid is converted to cytidine diphosphate diacylglycerol (CDP-DAG) by the condensation of cytidine triphosphate with phosphatidic acid. This reaction is catalyzed by CDP-DAG synthase and is freely reversible. The second step involves the action of the microsomal enzyme PI synthase to catalyze the condensation of *myo*-inositol with the CDP-DAG to form PI. This reaction requires Mg^{2+} as a cofactor and is selective for 1-steroyl,2-arachidonyl-glycerol-3-phosphate, a specific phosphatidic acid, thereby accounting for accumulation of this molecular species of PI within cells. *See* PHOSPHOLIPID.

Cell signaling. The *myo*-inositol moiety of PI has multiple sites which are sequentially phosphorylated by a family of specific kinases. The most important phosphorylated form of PI is phosphatidylinositol-4,5-bisphosphate (PIP_2). This form is localized inside the cell in two different sites: the inner leaflet of the plasma membrane and the nuclear membrane (**Fig. 2**).

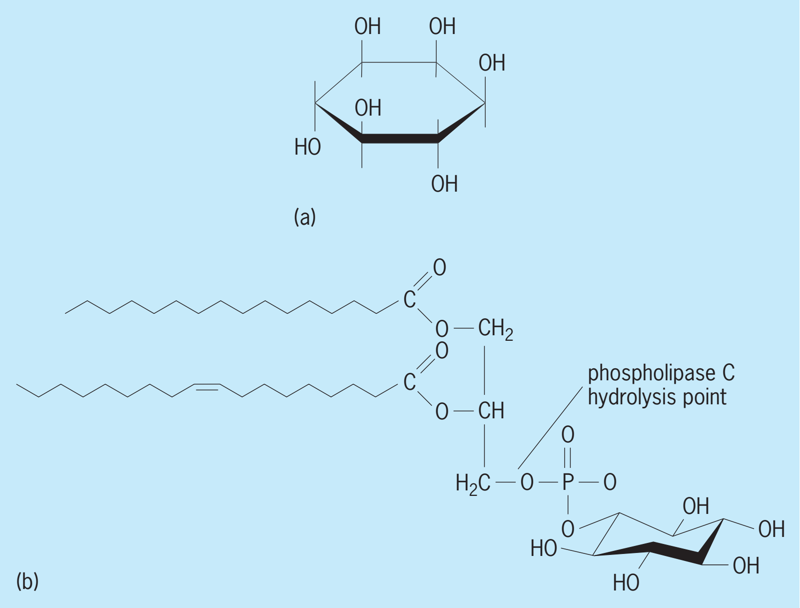


Fig. 1. Chemical structure of (*a*) D-*myo*-inositol and (*b*) phosphatidylinositol.

Cell signaling is turned on by the binding of various ligands to one of three different membrane receptor families. Upon ligand binding, phospholipase C (PLC) is activated, thereby initiating a potent intracellular signaling cascade. PLCs hydrolyze the PIP_2, releasing 1,4,5-inositol trisphosphate (IP_3) and diacylglycerol (DAG). These molecules potentiate the signal by their own downstream signaling events. IP_3 binds to specific intracellular receptors, releasing calcium from intracellular stores. DAG and the released Ca^{2+} then activate the protein kinase C (PKC) family, which then phosphorylate multiple different proteins, eliciting a variety of cellular responses. DAG lipase hydrolyzes the first fatty acid off the released DAG, making 2-arachidonyl glycerol (2-AG), a potent cannabinoid ligand that can act outside the cell (Fig. 2). Hence, the activation of a given cell surface receptor has multiple downstream effects, resulting in an amplification of the initial signal.

For every event that turns a signaling cascade on, there are a number of key steps that turn a signaling cascade off. The process described above is tightly regulated because of the tremendous amplification of the signal. Either IP_3 is phosphorylated to IP_4 by PI3 kinase or a phosphate is hydrolyzed by specific phosphatases, forming IP_2. Both of these inositol phosphates are inactive at the intracellular IP_3 receptors, so calcium release is stopped. Calcium is removed from the cell by intracellular calcium pumps and is returned primarily to the endoplasmic reticulum (ER). Eventually, through the sequential action of a family of phosphatases, the inositol phosphates have all of the phosphates removed, reforming *myo*-inositol. DAG is rapidly phosphorylated by DAG kinase to form phosphatidic acid. This phosphatidic acid and the *myo*-inositol are reused to form PI. This sequential set of events is commonly referred to as the PI cycle. *See* CALCIUM METABOLISM; SIGNAL TRANSDUCTION.

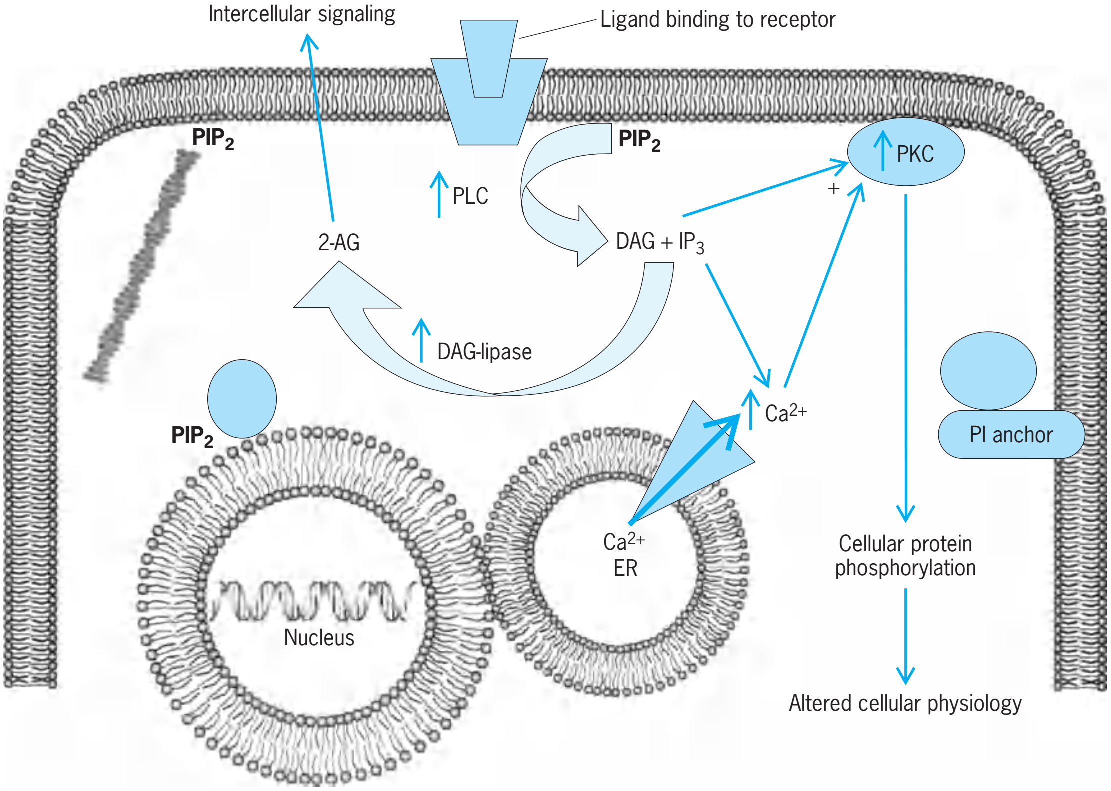


Fig. 2. **Process of cell signaling.**

Other functions. Another important function of phosphatidylinositols is the anchoring of membrane proteins through three distinct mechanisms.

(1) PI is glycosylated with a number of sugar moieties, including glucosamine, to form a polysaccharide chain on the PI. This glycosyl-PI is found in the plasma membrane inner leaflet, and free amine reacts with a number of proteins to anchor the proteins to the membrane.

(2) PIP_2 interacts directly with cellular proteins containing a certain sequence of amino acids called the pleckstrin homology domain. Recent discoveries demonstrate that this form of interaction is important in lipid-mediated signaling in the nuclear membrane, which aids in regulating gene expression.

(3) PI3 kinase phosphorylates the 3 position of the *myo*-inositol ring of PI, resulting in a family of PI's that have a unique function in cells. These PI3P's interact with a number of different proteins within the cell to promote membrane trafficking, cytoskeleton maintenance, PKC regulation, and cell survival. This includes regulation of important cellular processes such as endocytosis and exocytosis. *See* CELL MEMBRANES.

Eric J. Murphy

Bibliography. M. A. J. Ferguson and A. F. Williams, *Annual Review of Biochemistry*, 1988; P. W. Majerus, *Annual Review of Biochemistry*, 1992; C. K. Matthews and K. E. van Holde, *Biochemistry*, 2d ed., 1996; W. Pigman and D. Horton (eds.), *The Carbohydrates IA*, 1972; G. J. Siegel et al. (eds.), *Basic Neurochemistry: Molecular, Cellular, and Medical Aspects*, 6th ed., 1999; D. E. Vance and J. E. Vance (eds.), *Biochemistry of Lipids, Lipoproteins, and Membranes*, 4th ed., 2002.

Insect control, biological

The use of parasitoids, predators, and pathogens to reduce injurious pest insect populations and consequently the damage they cause. Viruses and bacteria are the most commonly used pathogens, but fungi, protozoa, and nematodes may also be important biological control agents. *See* PARASITOLOGY.

Ecological foundations. Three ecological assumptions underlie biological control. First, natural enemies are among the prime factors responsible for the regulation, or control, of pest populations. Second, the influence exerted by parasitoids, predators, and pathogens is density-dependent. Density dependence means that the killing power of the natural enemy increases as the prey or host density increases. Conversely, the mortality induced by density-dependent natural enemies decreases as host density increases. Both of these assumptions have become the subject of debate, and scientists have proposed alternative ecological assumptions that do not require density-dependent responses or stable interactions between natural enemy and prey or host. In the dominant, or classical, form of biological control the third assumption is found: when an insect species escapes into a new area without its natural enemies, it reaches outbreak levels and becomes a pest. Biological-control practitioners believe regulation can be reestablished by importing the natural enemies of the pest from its area of origin.

Classical approach. In classical biological control, all efforts are typically directed toward establishing the natural enemies that were left behind in the area of origin. Such an approach demands the verification of the area of origin, foreign exploration for natural enemies in the area of origin, importation of those natural enemies, quarantine, mass rearing, and release of the biological control agents. Quarantine procedures ensure that no natural enemy of the biological control agent selected—including a hyperparasitoid, which is a parasitoid of a parasitoid—is accidentally released. Mass rearing attempts to produce for release the largest number of vigorous and competitive biological control agents possible.

Classical biological control is by far the most frequently used form of biological control, assuming one excludes the use of resistant plant varieties as biological control. The Commonwealth Agricultural Bureau's International Institute of Biological Control has recorded 860 successful establishments of 393 species of parasitoids against 274 pest insects in 99 countries. In 216 cases, parasitoids by themselves or aided by predators achieved complete or satisfactory pest suppression, and an additional 52 cases are considered to have provided a useful reduction.

Conservation. Conservation involves manipulation of the environment in order to favor survival, reproduction, or any other aspect of the natural enemy's biology that affects its function as a biological control agent. Conservation can be undertaken regardless of whether the biological control agent is introduced or native, and generally involves tactics such as maintaining floral food sources, spraying sugar or yeast-sugar solutions for parasitoids and invertebrate predators, or providing nest boxes for vertebrate or invertebrate predators. Modification of agronomic practices or the timing of activities such as tilling or harvest residue destruction is another approach that may enhance the survival of natural enemies. For example, spraying insecticides in strips rather than across entire fields may provide an important degree of protection for natural enemies. Some natural enemies may become established but fail to spread beyond the area of release. That problem often may be solved by redistributing the biological control agent—collecting samples of the agent and releasing them at a distant, appropriate site. *See* INSECTICIDE.

Trends. Aspects of research on and application of biological control may provide new or improved approaches. Successes in the development of pesticide-resistant predatory mites suggest that the improvement of biological control agents through selection, hybridization, or genetic engineering techniques may play an important role. A major strategy for control of pest insects may involve the use of genetic engineering to introduce traits into natural enemies that enhance their performance, or mortality-causing traits of natural enemies, such as insect pathogens, into plants. Corn, tobacco, tomato, potato, and cotton have already been genetically engineered to incorporate the gene that regulates the production of the toxin normally produced by the bacterium *Bacillus thuringiensis*. Insect parasitoids and pathogens such as viruses are also the subject of genetic engineering research, which has attempted to clone and characterize the DNA encoding genes responsible for their toxicity or for characters that determine survival in the environment. *See* BREEDING (PLANT); GENETIC ENGINEERING.

Pedro Barbosa

Bibliography. M. A. Hoy and D. C. Herzog, *Biological Control in Agricultural IPM Systems*, 1985; G. C. Papavizas, *Biological Control in Crop Protection*, 1981; J. Waage and D. Greathead, *Insect Parasitoids*, 1986.

Insect diseases

A biological discipline embracing the general principles of pathology (disease in the broadest sense) as applied to insects. It refers to human observations and actions concerning the cause, symptomatology, gross pathology, histopathology, pathogenesis, and epizootiology of the diseases of insects; it is concerned with whatever can go wrong or become abnormal in an insect. A diseased insect may be suffering from an infectious disease caused by a microorganism or a noninfectious disease, such as metabolic disturbance, a genetic abnormality, a fetal malformation, a nutritional deficiency, a physical or chemical injury, or an injury caused by parasites or predators.

Insect pathology draws upon and contributes to the general field of microbiology and provides understanding of certain of the biological relationships existing between insects and microorganisms not pathogenic to them. Insect pathology finds applications in agriculture, medicine, and biology generally. Microbial control, the use of microorganisms in biological control, is one area of applied insect pathology. Microorganisms are introduced to control insect pests for the protection of humans and other animals and agricultural crops. However, the suppression of disease in beneficial insects, such as the silkworm, honeybee, and ladybird beetle, is also of significant practical importance. *See* INSECT CONTROL, BIOLOGICAL.

Pathogenesis and epizootiology. Many factors contribute to the condition or state described as a disease. It is important when diagnosing a disease to understand its pathogenesis, that is, how the pathology develops and how the individual host responds. Identifying the infectious microorganism or physical, chemical, or genetic factors affecting the insect is only a part of the knowledge needed for determining pathogenesis. Distinguishing one disease from another is as basically important in insect pathology as it is in the study of disease in other forms of life. One must know what is abnormal about or has killed an insect before the disease can be properly studied, controlled, or suppressed, used as a microbial control measure, its potentialities for natural spread determined, or its role in the ecology of an insect species known. *See* POPULATION ECOLOGY.

An important facet of insect pathology is concerned with the epizootiology of diseases, which involves the study of disease in populations of insects rather than in individual animals. Any epizootic affecting an insect population is concerned with three primary natural entities: the infectious agent, the insect host, and the environment. Each of these factors has certain attributes that, when properly related to the attributes of the others, play their appropriate role in determining the initiation, rise, and decline of an epizootic. Knowledge concerning the nature of epizootics is extremely important in insect ecology generally and in an understanding of insect-microbe ecosystems. The degree to which disease-producing microorganisms are dependent upon the density of the host population, the susceptibility of the host insect to the disease, the influence of weather and other environmental conditions on epizootics, the mode of transmission of insect pathogens, the success of the microorganism remaining alive, and other factors are important aspects of the epizootiology of insect disease. In the microbial control of an injurious insect, humans can initiate and increase the rapid rate of development of an epizootic by controlling the amount, method of distribution, and time of introduction of an effective microorganism.

Infectious diseases. While the noninfectious diseases and abnormalities of insects are very important and of numerous types, most of the activity in insect pathology has been with the infectious diseases. The infectious agents responsible for diseases in insects belong to the same major groups as those that cause such diseases in other organisms: the bacteria, fungi, viruses, rickettsiae, protozoa, and nematodes. In general, however, insects are not normally susceptible to those particular microorganisms that cause diseases of man, other animals, and plants. Moreover, most of the microorganisms that cause fatal diseases in insects are harmless to plants and to higher animals. Insects often can resist infection by humoral and cellular types of immunity; acquired humoral immunity is manifested in the hemolymph. However, it appears that the antibody immunity characteristic of higher animals does not occur in insects.

A convenient way to consider the various infections of insects is on the basis of the etiologic agent, as will be done in the sections to follow.

Bacterial Diseases

The characteristics of a bacterial disease in an insect may vary, but in general the following may be observed. As the disease develops, the insect becomes less active, has a smaller appetite, and discharges fluids from the mouth and anus. Infection may begin as a dysenteric condition with an accompanying diarrhea, but in most instances the invading bacterium eventually enters the body cavity of the insect and causes a septicemia that terminates in the death of the host. Following death, the insect's body usually darkens to brown or black as melanin pigment forms in the tissues. This is especially true of larvae and pupae in which the disintegration takes place rapidly.

The freshly dead insect is usually soft and flaccid. Internal tissues may disintegrate to a viscid consistency, sometimes accompanied by odor, but ordinarily they do not "melt," or liquefy, as do insects dying of certain virus infections. The cadaver usually dries and becomes shriveled, the integument remaining intact. Microscopic examinations of smears or of histological sections of an insect dead or dying of a bacterial disease usually disclose large numbers of the causative bacterium. These bacteria, if they are pathogens, must be differentiated from similar-appearing saprophytes that are normally in the insect intestine and that may flourish in the tissues of the dead insect. In some cases the bacterial pathogen may not be present in large numbers if it has killed the insect primarily through the production of toxins.

Spore-forming bacteria, which remain dormant for long periods of time if necessary and which are resistant to destructive environmental factors, have increased the probability, when compared to other bacteria, of contributing to epizootics with or without human manipulation.

Foulbrood. The first bacterial diseases of insects to receive concentrated study were flacherie of the silkworm, which Louis Pasteur considered at the time he was studying the protozoan disease pébrine of this same insect, and European foulbrood of the honeybee, which was studied by F. R. Cheshire and W. W. Cheyne in 1885. Later it was found that, in addition to European foulbrood, caused by a group of non-spore-forming bacteria, American foulbrood caused by the spore-forming bacterium *Bacillus larvae* also afflicted this beneficial insect. After the role of these bacteria in causing diseases of larval bees (hence the term foulbrood) was elucidated, it was possible to institute measures of control; however, in some areas these and other bee diseases still cause losses to beekeepers.

Milky diseases. In addition to the diseases of bees, two of the most important bacterial diseases of insects are the so-called milky diseases, which affect grubs (larvae) of the Japanese beetle (*Popillia japonica*) and are caused by *Bacillus popilliae* and *B. lentimorbus*; and the disease of numerous lepidopterous larvae caused by *B. thuringiensis*. All of these bacteria are sporeformers and are effective in the control of certain insect pests. Infection of Japanese beetle grubs occurs through ingestion of the bacterial spore, which germinates, penetrates the insect's intestinal wall, and develops in the body cavity, where the bacteria sporulate. The blood of the infected grub assumes the milky-white appearance that gives the disease its common name. Eventually the insect dies and decomposes, thus providing new spores which may be consumed by other grubs. In the eastern United States a dust containing milky-disease spores is used, at times in conjunction with chemicals and insect parasites, to control the Japanese beetle in lawns and turf. Distribution of this material has been carried on by the U.S. Department of Agriculture, which did much of the original work on the disease. Products containing the spores are commercially available.

Crystalliferous bacteria. *Bacillus thuringiensis* is one of the most intensively studied bacterial pathogens of insects because it has proved to be an effective microbial control agent against a number of insect pests, such as the alfalfa caterpillar, larvae of the cabbage butterfly, and larval forms of certain mosquitoes, black flies, and beetles. A number of closely related species of this bacillus are known to cause diseases of beneficial insects, such as silkworms. Several different strains or varieties of *B. thuringiensis* are known. Certain types affect only Lepidoptera, others only Diptera or Coleoptera. These bacteria are characterized by the formation of a protein crystal in the sporangium at the time of spore formation. The protein material comprising the crystal is highly toxic for certain insects, but apparently harmless for other forms of life. The crystal usually appears bipyramidal but may occur in other forms. Ordinarily each sporangium contains but one crystal; the crystal is freed from the cell wall with the spore and appears to persist indefinitely.

Bacillus thuringiensis and related species also produce a heat-stable, water-soluble substance, distinct from the crystal and from lecithinase, that is toxic for certain insects when injected, and to larval Diptera when they are exposed to it in a rearing medium.

Preparations containing the spores and crystals of *B. thuringiensis* or its varieties may be sprayed or dusted on crops or into aquatic habitats to destroy lepidopterous, dipterous, or coleopterous pests. Commercial products containing the bacillus are available. One of the primary advantages of this type of control is that this and other entomophilic microorganisms are nontoxic and noninfectious for plants and higher animals.

Bacillus sphaericus complex. Approximately 70 isolates of *Bacillus sphaericus* are described, from soil, mud, and vertebrate and invertebrate animals. Eight of these isolates release substances from the walls of spores in the alimentary tract of larval forms of mosquitoes. They are toxic for the insects and apparently harmless to all other forms of life. The bacteria can reproduce in the dead insect, as well as use other sources in habitats of mosquito larvae to provide additional spores.

Fungal Diseases

There are a large number of species of different groups of fungi that are pathogenic for insects. Their development on and in susceptible hosts depends upon the presence of adequate moisture to provide for the germination of conidia and spores, or on the presence of another invertebrate as an alternative host for development of the form of the fungus infectious for the insect. Whereas bacteria, viruses, and protozoa infect their insect hosts by way of the mouth and digestive tract, most entomophilic fungi enter their hosts directly through the integument or body wall.

Many of the fungi associated with insects are not true pathogens, or are pathogenic only under certain conditions. These saprophytic species frequently develop in or on insects that have died of other causes, since the conidia of the fungi may be closely associated with the living healthy insect. Some fungi appear to be parasitic but not lethal in their association with insects; thus the large order Laboulbeniales (Ascomycetes) is made up of species that live primarily on the external surfaces of insects. The genus *Septobasidium* consists of species associated with scale insects which, although some are parasitized, live beneath the stromata of the fungi. Some fungi, such as the ambrosia fungi of certain termites and wood-boring beetles, are definitely mutualistic in their relationship with insects. *See* ASCOMYCOTA; BASIDIOMYCOTA.

Fungal development. Once the germ tube from a conidium is within the body cavity of an insect, the threadlike hypha of the fungus proliferates, invades the tissues, and fills the body of the insect. In most cases the fungus then develops conidiophores to the exterior of the host body, where conidia develop to provide a means for initiating development of the fungus in new hosts. The infected insect usually assumes a dried mummylike appearance, often is covered with a conidia-bearing mycelium, or can contain resting spores which enable the fungus to survive periods of adverse environmental conditions or absence of a suitable host. Toxins may be produced by the fungus during the initial period of hyphal growth.

Entomophilic fungi. Some entomophilic fungi occur in the orders Mucorales, Blastocladiales, and Chytridiales, but the most important fungi of the class Zygomycetes belong to the order Entomophthorales. In this order the single family Entomophthoraceae is made up of several genera, of which *Entomophthora* and *Massospora* are composed primarily of entomophilic species. One of the most commonly known species is *E. muscae*, which attacks the housefly. A large number of *Entomophthora* species are known. They infect many species of Diptera, Orthoptera, Lepidoptera, and other insects.

Infected houseflies are frequently found indoors attached to the walls and ceilings of buildings in a lifelike position. Close inspection of flies killed by the fungus usually reveals on the wall or windowpane a distinct halo of discharged spores (conidia) encircling the insects. This halo results from the fact that, after the conidiophores have emerged through the integument of the insect, conidia are formed which are discharged violently into the air from the terminal portion of the conidiophores, thus forming a ring of conidia about the dead insect. When the discharged conidium comes into contact with a susceptible insect in the presence of adequate moisture, it begins to germinate, sending out a conidial hypha (germ tube) that penetrates through the integument into the body cavity.

Blastocladiales. The entomophilic Blastocladiales, class Chytridiomycetes, are confined largely to one group, the family Coelomomycetaceae, which parasitizes mainly mosquito larvae. About 50 species are known, all in the genus *Coelomomyces*. Investigations of the biology and life cycle of *Coelomomyces* species have demonstrated the requirement for the presence of invertebrate animals, recognized as copepods, as alternative hosts for the fungus pathogen. Infectious forms from the larvae of mosquitoes infect the copepod, and the infectious agents from the copepod initiate the fatal infection in the mosquitoes. Certain regions of the body cavity or virtually the entire hemocoel of the mosquito may become filled with the spores and mycelium of the fungus.

Hypocreales. Species of the genus *Cordyceps*, an ascomycete of the order Hypocreales, are among the first recorded entomophilic fungi. They are frequently of large size and very colorful in appearance. About 250 species are known. They are cosmopolitan in distribution and occur on representatives of several orders of insects, principally Hemiptera, Diptera, Lepidoptera, Hymenoptera, and Coleoptera. In members of the genus *Cordyceps*, the stroma arises from a sclerotium formed within the body of the infected insect. At the end of this stroma, or stem, is a fertile portion, the "head," which may be brightly colored. The life history of most *Cordyceps* is similar to that of entomophilic fungi in general.

Muscardine. One of the most important groups of entomophilic fungi is responsible for muscardine diseases, in which the fungus emerges from the body of the insect and covers the animal with a characteristic fungus mat resembling, in a way, a French bonbon or candy mint. The word muscardine was first used for white muscardine, a well-known disease of the silkworm caused by the fungus *Beauveria bassiana*. It has also been used in reference to green muscardine, a disease of the European corn borer, the wheat cockchafer, and other insects caused by *Metarhizium anisopliae*. The mycelium from conidia can enter the respiratory siphon of mosquitoes at the water surface and grow, resulting in the death of larvae. These fungi belong to the Fungi Imperfecti. *Beauveria bassiana* and closely related species have been identified from a large number of other insects. In certain situations the fungus has shown some promise as a means of control of pest insects. The fungi produce toxin in the host, accelerating the death of the insect. Among their best known hosts, in addition to the silkworm, are the European corn borer, codling moth, chinch bug, and the alfalfa weevil. *See* DEUTEROMYCOTA.

Viral Diseases

The insect species infected by viruses are principally in the order Lepidoptera, although a few Hymenoptera and Diptera and one or two Coleoptera and Neuroptera are known to suffer infection with these agents. Only the immature stages, larva and pupa, are killed by virus. Adults may carry the virus, but they are not killed by it. In general, many insect viruses exhibit a fairly high degree of host specificity, but numerous instances of cross infectivity have been recorded. *See* VIRAL INCLUSION BODIES; VIRUS.

Polyhedroses. The polyhedroses are characterized by the formation of polyhedron-shaped inclusion bodies in the infected tissues of the host. These proteinaceous inclusion bodies contain the virus particles embedded in their matrices. In the nuclear polyhedroses the virus multiplies in the nuclei of the infected cells, and in the cytoplasmic polyhedroses the virus multiplies in the cytoplasm of such cells. As far as is known, viruses causing nuclear polyhedroses are rod-shaped and contain deoxyribonucleic acid (DNA); those causing cytoplasmic polyhedroses are more or less spherical and contain ribonucleic acid (RNA).

Larvae infected with the virus of a nuclear polyhedrosis usually show few distinctive symptoms until a few hours before death. The incubation period varies in different host species between 5 and 20 days; usually it covers a period of about a week. In some species of insects, the infected larvae may cease feeding, become somewhat sluggish in movement, and become yellowish or pale in color. They may swell slightly, then become limp and flaccid. The tissues most prominently affected are the fat-body cells, epidermal cells, blood cells, and the cells of the tracheal matrix. Shortly before and after death the integument is very fragile and easily ruptured, emitting the liquefied contents, which are filled with disintegrating tissue and polyhedra. Polyhedra increase in size and number in the nucleus of the cell until the nuclear and cell membranes are disrupted. The dead larvae are usually found hanging by their prolegs from the host plant or other support. Eventually, they may burst or dry to a dark-brown or black cadaver.

The polyhedra are insoluble in water, alcohol, ether, and acetone but soluble in weak acids and alkalies. They range in size from 0.5 to 15 micrometers. If the polyhedra are treated properly with dilute alkali, the rod-shaped particles within them can be demonstrated. The size of the virus rods ranges from about 20 to 50 nanometers in width and from about 200 to 350 nm in length. The virus rods may occur singly or in bundles of up to as many as eight rods.

Production in living larval Lepidoptera of nuclear polyhedroses in large quantities to control agriculture and forest insects is an accepted practice throughout the world. In the United States, the Environmental Protection Agency has approved virus products for use against the cotton bollworm, the gypsy moth, and the tussock moth.

The cytoplasmic polyhedroses are, in general, not so fulminating as the nuclear polyhedroses. The host integument does not become fragile, as with the nuclear polyhedroses. In the cases known, the infection is limited largely to the midgut epithelium, where the polyhedra are formed in the cytoplasm of the cells. The polyhedra are morphologically similar to those present in nuclear polyhedroses. The virus particles, in contrast, tend to be spherical in shape, about

20–70 nm in diameter, and near the surface of the polyhedron. One of the best-known cytoplasmic polyhedroses is that of the silkworm, which also suffers from a nuclear polyhedrosis. *See* DEOXYRIBONUCLEIC ACID (DNA); RIBONUCLEIC ACID (RNA).

Granuloses. The granuloses have so far been found only in larvae of Lepidoptera. They are characterized by the accumulation of small granular inclusion bodies, called capsules, in the infected cells. Each capsule contains a single virus rod composed of DNA. Compared to the nuclear polyhedroses, the symptoms associated with the granuloses are more subdued; the diseased larvae usually become less active, somewhat flaccid, and assume a pallid or whitish translucent aspect. The integument does not become fragile. The period from infection to death ranges from 6 to 20 days. The inclusion bodies, or capsules, are about 200 by 500 nm in size, and the virus rod ranges, depending on the virus, from 40 to 80 nm in width and from 200 to slightly over 300 nm in length.

Other viruses. Among the insect viruses so far discovered that are not associated with an inclusion of any kind are those which infect certain Lepidoptera (army worms), Hymenoptera (honeybees), Coleoptera (beetle grubs), and Diptera (mosquitoes, crane flies). A number of the noninclusion viruses exhibit a unique physical characteristic when concentrated in large quantity. Centrifugate pellets of the virus particles appear orange or amber in color by transmitted light and iridescent or turquoise by reflected light. Another peculiarity is the fact that an exceptionally large amount of virus is produced in the host tissues. A noninclusion virus called "sigma virus" is known to occur in *Drosophila*. This virus renders infected flies sensitive to carbon dioxide. Among acarines, noninclusion viruses have been found in at least two species of mites.

There is considerable evidence that insect viruses frequently remain occult and cause latent infections in their hosts. Some authorities believe that a virus can survive in a host, even for several generations, without causing recognizable symptoms, but that under the influence of certain stresses or incitants, such as cold, crowding, and certain chemicals, can be triggered into causing an active or frank infection.

Rickettsial Diseases

A number of rickettsiae or rickettsialike organisms are capable of causing frank infections in insects. Three or four have been found in Coleoptera and one in Diptera. These infections appear to be slow to develop and kill their hosts in 1–4 months. The fat body of grubs assumes a characteristic bluish-green coloration, and in some host species crystalline bodies occur in association with the rickettsie.

Protozoan Diseases

Protozoan infections in insects may be benign and cause little morbidity or mortality, but many are severe and fatal. Some appear in epizootic proportions, while others may be represented by only a local invasion of tissue by the protozoon. The success of reproducing populations of insects can be seriously affected by destruction of gonads without death of the host. Generally, protozoan diseases are relatively slow in developing, and frequently become somewhat chronic in nature; but some sporozoan infections develop rapidly and kill the insect host within a short time. Infected insects may show few, if any, external signs of infection, or may be stunted in growth and development or not undergo metamorphosis, or show a change in translucency and color. Usually they become opaque and whitish because of the accumulation of spores or cysts in the internal tissues or fluids. Further, infected insects may exhibit a loss of appetite, abnormality of movement, and may remain moribund for long periods prior to death.

Only occasionally do flagellates cause acutal disease in insects. A few amebas cause disease. There is an important amebic disease in the honeybee and another in grasshoppers. In both of these maladies the insect's Malpighian tubes are especially involved. Several species of Ciliata have been found infecting insects; other are frequently found adhering to the exterior of insect larvae and pupae living in aquatic environments. Perhaps the best known are ciliates associated with mosquitoes. *See* AMOEBIDA; CILIATEA; MASTIGOPHORA; TRYPANOSOMATIDAE.

Sporozoa. The most important protozoan diseases in insects are caused by Sporozoa, of which the Microsporida as a group are the most serious. Many Gregarinida are found in insects. Most of the so-called eugregarines from the alimentary canal are not pathogenic for their hosts, but the schizogregarines found in the body cavity may be highly virulent. A number of Coccidia are also pathogenic for insects. *See* SPOROZOA.

Microsporida. Of the infections caused by members of the order Microsporida, perhaps the most famous is pébrine of the silkworm, caused by *Nosema bombycis*. Another well-known infection is nosema disease of the honeybee, caused by *Nosema apis*. Intensive studies of these two diseases have revealed a great deal about the Microsporida. Microsporidan infections have been found in species of at least 14 orders of Insecta. The most common insect hosts are Diptera, Lepidoptera, Coleoptera, and Ephemerida. Some of these are rather specific as to their insect hosts, while others are capable of infecting insects in different genera or even in different orders.

Spore stage. The most commonly seen form of any Microsporida is the spore stage, the characteristics of which are important in distinguishing the different species. The spore serves as the resistant stage of the organism and is able to tide the pathogen over periods of unfavorable environmental conditions and during the period between the change of host individuals. The average microsporidan spore is 3–6 μm long and 1–3 μm broad: some species exhibit a range of spore sizes. The shape of the spore is usually oval-spheroidal, although it may be spherical or bacilliform. The spore consists of a spore membrane or covering surrounding a sporoplasm and a polar filament coiled within the spore. The polar filament is capable of extrusion as a very fine, long thread that may be several hundred micrometers in length in some species.

Microsporidan development. Although the life cycles of Microsporida vary from species to species, in general, soon after the spore is ingested by a susceptible host, the polar filament is extruded. A successful infection is initiated when the tip of the polar filament deposits sporoplasm from the spore either within or between the epithelial cells of the insect's intestine. The parasite, now called a planont, may remain in the intestinal cell or pass into the body cavity, where in either location it multiplies by binary fission. Various tissues of the body can be invaded, and the developing protozoon is now designated as a schizont or meront. In their intracellular location the spherical-to-oval meronts divide actively by fission, budding, or multiple division, filling the cytoplasm of the host cell. This period of schizogony ends in the formation of sporonts, each of which produces one sporoblast or many, depending on the genus of the microsporidian. Each sporoblast becomes a spore. This is the sporogony part of the life cycle. The entire life cycle is usually completed in about 4 days. Insects are usually infected with Microsporida through the mouth. However, some of these protozoa are transmitted from one generation to the next by way of the eggs.

Evidence is provided in India and the United States in support of the utility of *Nosema algerae* to infect larvae and for the microsporidia to interfere with the development of the malaria parasite in adult female mosquitoes. *Nosema locustae* infects grasshoppers and is a microsporidian which can be mixed with a bait to increase the consumption of the parasite by grasshoppers and consequently their mortality.

Nematode Diseases

Those roundworms for which insects serve as primary hosts are included in the classes Nematoda and Nematomorpha. Some members of the class Acanthocephala spend their larval stage in an insect host and the adult stage in mammals. More than 1000 species of roundworms have been reported from insects; most of these belong to Nematoda that frequently kill or seriously harm their hosts. At least 16 orders of insects are involved as hosts to nematodes. *See* ACANTHOCEPHALA; NEMATA (NEMATODA); NEMATOMORPHA.

All nematodes are similar in their general external appearance. Their elongate, unsegmented body has little or no variation in diameter, although it does taper toward one or both ends. There are three main stages in the developmental cycle of most nematodes: egg; juvenile, or larval, which includes four growth stages; and adult. Usually the young larvae spend a short period as free-living organisms, frequently in an aquatic environment.

The biological relations between nematodes and insects vary all the way from mere fortuitous association to obligate parasitism. For convenience these relationships may be separated into three general groups: (1) nematodes that live in the alimentary tract of the insect in a more or less commensal association; (2) nematodes that combine saprophagous and parasitic habits, that is, semiparasites; and (3) nematodes that parasitize the body cavity or tissues of their hosts, called obligate parasites. Examples of the first relationship are found in the alimentary tracts of crickets and cockroaches. These nematodes cause little, if any, harm to the insect. Of the second group, one is *Neoaplectana glasseri*, which infects the Japanese beetle. *Romanomermis culicivorax* is an obligate parasite which has shown some promise as a biological control agent against larval mosquitoes. Several commercial attempts in the United States to develop economical production of nematodes have not been successful. Among the best-known obligate parasites are the Mermithidae for safe and effective use in terrestrial habitats.

Noninfectious Diseases

Although most of the present activity in insect pathology appears to be concerned with the infectious diseases, increasing attention is paid to the noninfectious diseases of insects. These noninfectious diseases and abnormalities may be grouped as follows: (1) mechanical injuries such as traumata, bruises, and torn tissues; (2) injuries caused by physical agents such as burning, freezing, and drought; (3) injuries caused by poisons or chemical agents such as insecticides; (4) injuries caused by parasitization or infestation by other insects or arachnids; (5) diseases caused by a nutritional disturbance or a deficiency of proper nutrients or vitamins; (6) diseases caused by deranged physiology and metabolism; (7) inherited abnormal conditions or genetic diseases; (8) congenital anomalies and malformations; (9) certain tumors and neoplasms; and (10) disturbances in development and in regenerative capacity of tissues. *See* INSECT PHYSIOLOGY; INSECTA; INSECTICIDE.

John D. Briggs

Bibliography. N. E. Beckage et al., *Parasites and Pathogens of Insects*, 1993; A. E. Hajek and R. J. St. Leger, Interactions between fungal pathogens and insect hosts, *Annu. Rev. Entomol.*, 39:293–322, 1994; H. K. Kaya and R. Gaugler, Entomopathogenic nematodes, *Annu. Rev. Entomol.*, 38:181–206, 1993; P. R. Shewry and S. Gutteridge (eds.), *Plant Protein Engineering*, 1992; Y. Tanada and H. K. Kaya, *Insect Pathology*, 1997.

Insect physiology

The study of the functional properties of insect tissues and organs. Insects are one of the most diverse groups of animals on Earth. They have evolved adaptations that allow them to live in a variety of ecosystems. The basic body plan comprises a hard exoskeleton forming a protective covering over internal tissues. The interior of the exoskeleton is referred to as the hemocoel. Insects have an open circulatory system with hemolymph (blood) freely moving within the hemocoel and bathing all of the internal tissues. A basement membrane composed of a noncellular matrix of proteins, lipids, and carbohydrates separates the internal tissues from the hemolymph. The fat body is a specialized tissue within the hemocoel. It is usually a whitish tissue found throughout the body of the insect and is involved in storing fat reserves. It is also involved in general metabolism and regulates the composition of the hemolymph. The fat body is, in fact, the site of production for the major proteins found in the hemolymph. *See* INSECT PATHOLOGY; INSECTA.

Development and Growth

Insects undergo development in stages. In general, development is either hemimetabolous or holometabolous. Hemimetabolous insects have three life stages: the egg, larva (or nymph), and adult. Holometabolous insects have four life stages: the egg, larva, pupa, and adult. The larval stage or nymph of hemimetabolous insects is similar in appearance to the adult. Holometabolous larval stages do not resemble the adult stage, and a pupal stage is present. Holometabolous development is considered complete metamorphosis because transformation from larva to adult occurs in the pupal stage. Hemimetabolous development is referred to as incomplete metamorphosis because of the gradual change from an immature nymph to an adult. The life stages are characterized by progressive development in which the cuticle is shed. Typically, the larval stages have more than one stage and undergo successive molts to become larger. *See* METAMORPHOSIS.

Embryonic development takes place within an eggshell called the chorion. Fertilization takes place within the egg. In some cases, parthenogenesis can occur, in which development takes place without fertilization. Many species of ants, wasps, and bees produce males through parthenogenesis. However, in these insects the majority of individuals are female and are produced by fertilization. In these insects the females are diploid (having two complete chromosome pairs in a nucleus), and the males are haploid (having half of the diploid or full complement of chromosomes). The summer generations of aphids also reproduce without fertilization, but all are female. Males can be produced in the fall for sexual reproduction. Most insects lay their eggs externally, after which they undergo embryogenesis and hatch. In some insects the eggs are retained in the female until hatching, and the mother gives birth to live young; aphids, tsetse flies, and some cockroaches are examples.

Upon hatching, the larva starts to consume food and grow. When growth reaches a certain size, the insect must shed its old cuticle, and a new cuticle is formed. This process is called molting or ecdysis. The cuticle of insects is part of the hard exoskeleton that is characteristic of the arthropods. The cuticle is an acellular layer that is produced by the underlying cellular epidermal cells. It is composed of several layers that are made up of protein, chitin, catecholamines, and lipid. Chitin is a complex polymer composed of repeating units of *N*-acetylglucosamine linked together. The chitin and proteins are cross-linked together by catecholamines such as *N*-acetyldopamine, and this cross-linking creates the rigidity found within the cuticle. The cuticle can become quite hard, or it can remain soft and flexible. The different types of protein, amount of chitin, and catecholamines are thought to contribute to the varied types of cuticle that are found in the insects. Based on the chemical properties of cuticle proteins, as determined by structural analysis and molecular genetic analysis, it is now known that the more acidic proteins are found in flexible cuticle. Molecular techniques have also revealed that epidermal cells and hemocytes (cellular elements of the hemolymph) are the source of these proteins, and at what time during development particular proteins are produced and become incorporated into the cuticle.

Lipids are located on the outer surface of the cuticle. These form a barrier that is primarily responsible for protecting against water loss in terrestrial insects and against water gain in freshwater insects. In addition, the nonpolar lipids help the insect resist attack by microbes. Some of the lipids found on the surface are also used in chemical communication. Hydrocarbons especially are utilized by a variety of insects for communicating.

The old cuticle is shed at each molt in a process that can occur over several days. Prior to ecdysis, the epidermal cells begin to divide and become close-packed and columnar in form. The old cuticle then breaks away from the epidermal cells in a process called apolysis. This creates a space between the epidermal cells and the old cuticle. New cuticle is then produced on top of the epidermal cells, and a molting fluid is secreted into the subcuticular space. After a portion of the new cuticle is produced and hardened, the molting fluid is activated and digests away some of the old cuticle. Many of the products of digestion are absorbed back into the epidermal cells. The next step is ecdysis, which is the actual shedding of the old cuticle. The old cuticle splits open, and the insect draws itself out. The old cuticle is then called an exuviae. The new cuticle is expanded and then hardens over several hours. The hardening process is partly due to the cross-linking of proteins. When they first emerge, many insects have a white cuticle, which slowly hardens and changes color to black or brown. The peptide hormone bursicon regulates the hardening of the cuticle after an ecdysis.

The overall process of molting is under the regulation of several hormones. The main ones are juvenile hormone and ecdysone. The juvenile hormone is responsible for maintaining the juvenile or larval stage upon subsequent molts. The actual molting hormone is ecdysone. Juvenile hormone ($Cl_6H_{26}O_3$) has a sesquiterpene type of structure, and is produced by endocrine glands called corpora allata. Ecdysone is a steroid that is produced by the prothoracic glands. It is a prohormone (a hormone precursor) that is converted to the active hormone, 20-hydroxyecdysone, by epidermal or fat body cells. In general, it is the action of juvenile hormone that governs what type of cuticle and internal tissues will occur at the next molt. If juvenile hormone is present during a critical period of a larval stage, the next stage will be larval. For hemimetabolous insects, if juvenile hormone is absent during the last nymphal stage, the next molt will result in a change to the adult. Holometabolous insects have a pupal stage, and if juvenile hormone is absent during a critical period but present during a subsequent critical period, the next molt will result in a pupal stage. During the pupal stage, juvenile hormone is completely absent and the adult is formed within the pupal case. Ecdysone controls the molting process, and it is present before the molt to transform the tissues for the actual ecdysis.

A receptor protein for ecdysone has been identified through molecular genetic techniques. The receptor for ecdysone also binds to another protein called ultraspiracle. This complex of proteins bound to ecdysone is found within the nucleus and initiates gene transcription. Both ecdysone and juvenile hormone act by regulating the transcription of specific genes. The receptor for juvenile hormone has not yet been characterized. Other hormones are involved in regulating the complex process of actual ecdysis. These include the peptides eclosion hormone and ecdysis-triggering hormone. *See* ECDYSONE; ENDOCRINE SYSTEM (INVERTEBRATE); MOLTING (ARTHROPODA).

Diapause

Insects that live in temperate climates must be able to overwinter. In some cases, insects in the Northern Hemisphere migrate south for the winter. For example, the monarch butterfly overwinters in Mexico and southern California. However, most insects do not migrate and have developed a means to help them survive adverse conditions. This is called diapause, a physiological state into which the insect enters in advance of the actual onset of adverse environmental conditions. In most cases, diapause is also under genetic control. During diapause, insects become quiescent, metabolic rates decline, and the insects develop various physiological and biochemical adaptations to the cold environment.

The environmental conditions that induce diapause are primarily the photoperiod and temperature. Usually a critical day length induces diapause development. In temperate climates, diapause is triggered by shorter day length, which indicates that winter is approaching. Cooler temperatures act synergistically with photoperiod. Declining food availability is another indicator to some insects to enter diapause.

Diapause can occur in all of the different life stages of insects, including embryos in the egg stage. The silkmoth has an embryonic diapause that is determined by whether or not the mother underwent diapause. This insect has two generations per year, with the first generation diapausing as an egg. These eggs hatch and develop into adults that lay eggs that do not diapause. However, the second generation lays diapause eggs that survive the winter. Embryonic diapause in the silkmoth is under the control of a peptide hormone (diapause hormone).

Diapause can also occur in the larval, pupal, and adult stages. The same hormones that regulate metamorphosis (juvenile hormone and ecdysone) usually regulate these types of diapause. The cessation of feeding and an alteration in behavior to find a place to undergo diapause is characteristic of these stages. The stages also start producing cold-protective agents, usually alcohols such as glycerol that act as antifreezes. In insects that have an obligatory diapause, a certain period of cold exposure is required to break diapause and continue development. How an insect measures this cold period is probably coupled to the internal biological clock.

System Physiology

Insects, like all higher animals, have physiological mechanisms to support life. These functions are effected by the respiratory system for acquiring oxygen and releasing carbon dioxide, the muscular system for movement, the digestive system for acquiring nutrients, the circulatory system for movement of blood, the nervous system for perception of the environment and internal regulation, the reproductive system, as well as other systems.

Respiration. Most insects acquire oxygen and release carbon dioxide through a series of air-filled tubes called tracheae. The tracheae are connected to the cuticle and open externally through valves called spiracles. Insects can have as few as one pair or up to ten pairs of spiracles. The opening and closing of the spiracles in part regulates the airflow in the tracheae. Tracheae are composed of cuticle with underlying cells. They are large in diameter and supported by spiral thickenings of the cuticle called taenidia, which help to prevent collapse of the tracheae. Tracheae become progressively smaller as they integrate with the tissues, and are known as tracheoles. Almost every cell in the insect is associated with a tracheole. Diffusion accounts for the movement of oxygen within the tracheal system. Respiring cells acquire the oxygen, which creates a diffusion gradient. Carbon dioxide is released from the cells and can be converted first to bicarbonate for intracellular storage and buffering. When the bicarbonate concentration becomes too high, carbon dioxide is released as a gas and diffuses to the outside environment through the tracheae and spiracle system. Carbon dioxide causes the spiracle muscles to relax, resulting in opening of the spiracle.

In addition to the tracheae and tracheoles, insects have air sacs. These are characterized as large-diameter tracheae with reduced taenidia to allow for expansion and deflation. During periods of activity when an insect may need more oxygen, these air sacs expand and fill with air. To force air through the tracheal system, the spiracles close and the air sacs are forced to deflate. The rhythmic expansion and contraction of the abdomen along with the timing of spiracle opening and closing create active ventilation.

Aquatic insects also have a spiracle and trachea system. These insects, however, have adaptations to living under water. Some insects have a gill system that allows them to acquire oxygen dissolved in water, but the tracheal tubes are still gas-filled. Very few insects utilize hemoglobin for transport of oxygen. Diving insects acquire oxygen by periodically returning to the surface to replenish their air supply. In some cases, these insects create an air bubble that is maintained against the outer surface of the exoskeleton. This air bubble is held in place by hairs or under the wing, and is called a plastron. *See* RESPIRATION.

Nutrition and digestive system. Insects are the main competitors for food and fiber consumed by humans; indeed, they consume almost any type of food source. Various herbivorous insects eat almost every kind of plant material, either directly or by ingesting plant fluids. Many insects prey upon other animals; carnivore insects primarily consume other insects. Liquid blood is also consumed by mosquitoes, other biting flies, fleas, and some true bugs (of the order Hemiptera), which may make them transmitters of blood-borne pathogens. Larval stages of some insects, notably some Hymenoptera, are parasitic on other insects. The larvae feed on the living host but eventually kill it. These types of parasitic insects are called parasitoids.

Despite the varied dietary lifestyles of insects, nutritional requirements are similar. Insects require the 10 essential amino acids, vitamins, and choline. A protein source is usually required for development of eggs. Sugars are usually found in insect diets. Insects have a dietary requirement for cholesterol or another steroid, since they cannot make the cholesterol that is necessary for cell membranes and is the precursor for the steroid molting hormone ecdysone. Most insects also require polyunsaturated fatty acids for cell membrane dynamics and for biosynthesis of prostaglandins. However, some insects have the ability to make the diunsaturated fatty acid linoleic acid. Plants were thought to be the only source for this fatty acid until it was demonstrated that some insects could biosynthesize it.

Insects usually have a symbiotic relationship with microorganisms, which are found within the gut lumen and help the insect in digestion. In some cases, there are specialized cells called mycetocytes that house the symbionts. These cells are usually found within the hemocoel. The blood-feeding insects require the presence of the symbionts for survival, since they provide the vitamin B complex. These microorganisms can be passed on to the next generation by feeding to the young or through transovariole transmission, passing the symbiont to the developing egg.

The digestive tract of insects consists of a tube that runs from the mouth to the anus (**Fig. 1**). It is usually divided into the foregut, midgut, and hindgut. The foregut and hindgut are lined with cuticle. In most insects the midgut includes the peritrophic matrix, which is a chitin-protein complex that surrounds the food. Salivary glands release digestive enzymes and other factors in the mouth for extra-oral digestion and to begin the digestive process. The foregut takes in the food and in some cases can store food (in the crop). Most of the digestive enzymes are released from the midgut, and absorption of nutrients occurs in the ventriculus. The release of digestive wastes occurs in the hindgut. *See* DIGESTION (INVERTEBRATE).

Excretory system. The excretory system in insects involves the malpighian tubules, which produce the primary urine that flows into the digestive system and hindgut. These closed-end tubes connect to the gut at the intersection of the midgut and hindgut, and they are involved in secretion of nitrogenous and other waste products into the hindgut. The number of malpighian tubules varies from 2 to 200. They are usually free-moving within the hemocoel and clear

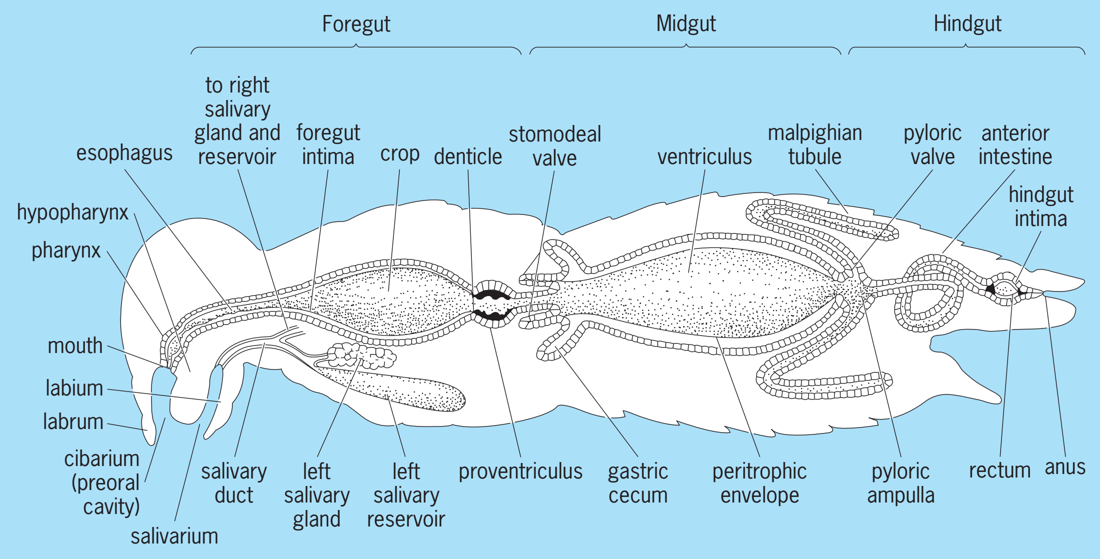


Fig. 1. Generalized insect digestive system. (*After W. S. Romoser and J. G. Stoffolano, Jr., The Science of Entomology, 4th ed., McGraw-Hill, 1998*)

the hemolymph of unwanted products. The primary urine formed in these tubules is water-based and contains nitrogenous waste products, usually in the form of uric acid for terrestrial insects and ammonia for aquatic insects. Other forms of nitrogenous waste are allatonin and urea. Other substances in the primary urine include ions (chloride, sodium, potassium), amino acids, and sugars. The ions are involved in providing the water flow through the malpighian tubules to the hindgut. Water is conserved in the hindgut and is absorbed back into the hemolymph. The hindgut also absorbs most of the amino acids and sugars and any ions that are in short supply. The feces consist of undigested food products and concentrated urine. *See* EXCRETION.

Water regulation. Water balance is maintained in insects by taking in water from the diet. The cuticle provides a very good barrier against water transport. The excretory system is also involved in water balance by regulating the amount of water lost in the feces. For example, if water needs to be conserved, the hindgut absorbs more water from the feces into the hemolymph. Two types of peptide hormones control water regulation in the excretory system. Diuretic hormones are released when the insect senses that too much water is present in the hemolymph. These act on the malpighian tubules to secrete more water and on the hindgut to inhibit the reabsorption of water. Antidiuretic hormones have been identified in a few insects, and these act primarily on the hindgut to increase the uptake of water from the feces. Some diuretic hormones have been identified; they are similar in sequence to the corticotropin-releasing factor of vertebrates. A receptor found on malpighian tubules of *Manduca sexta* (tobacco hornworm) was isolated and cloned. It possesses seven putative transmembrane domains, typical of the G-protein [GTP (guanosine 5′-triphosphate)–binding protein] coupled receptors. With the relatively recent sequencing of several insect genomes, other peptide hormones and their receptors have been identified. These diuretic hormones act on the malpighian tubules by stimulating the production of cyclic adenosine monophosphate (cAMP), which in turn regulates the ion transport machinery of the cells. Fluid secretion is stimulated by the active transport of ions and by a proton pump. The active transport of ions sets up a concentration gradient which water follows.

Circulation, hemolymph, and hemocytes. Insects have an open circulatory system. The exoskeleton forms the outer protective covering; internal tissues are found within the open hemocoel. The internal cavity is bathed in hemolymph or blood. Thus, all of the internal tissues are bathed in a hemolymph, with only a basement membrane separating the hemolymph from the tissues. Hemolymph flows backward and mixes within the insect by movement of the skeletal and visceral muscles. A dorsal vessel (essentially a hollow tube that runs the length of the insect) then circulates hemolymph forward again. This tube is composed of a one-cell-layer-thick heart that circulates hemolymph from the abdomen to the head. The heartbeat is produced by rhythmic contractions of the heart muscle. The heart is innervated by neurons, but since it will continue to beat in the absence of neural input the heartbeat is said to be myogenic. The heartbeat increases during activity, such as flight, and during the molting period. Several compounds affect the heartbeat, including biogenic amines such as octopamine and cardioacceleratory peptides. Special additional pulsatile organs are used to circulate hemolymph into the wings and antennae.

The hemolymph is composed of a variety of compounds. It can be colorless, or yellowish or green due to the presence of pigments. The most common sugar present in insect hemolymph is trehalose (a disaccharide of glucose units). Lipids are transported through the hemolymph in the form of diacylglycerol. Lipophorin is a protein carrier that holds the lipid and hydrocarbons for transport. A variety of proteins are present, most of which are made by the fat body. Insect hemolymph also contains high levels of amino acids, which are used to help maintain the proper osmotic pressure. Insects that are exposed to cold temperatures can also increase the levels of antifreeze alcohols, such as glycerol, in the hemolymph.

Hemocytes are the insect's blood cells. These are free-floating in the hemolymph, and are composed of several different types. Hemocytes are primarily involved in the insect immune response, which can be divided into cellular and humoral. Cellular responses involve the hemocyte first recognizing an invading pathogen. How this recognition occurs is not well understood but is most likely based on cell surface chemistry of the pathogen. The next response is the phagocytosis of pathogens that are smaller than the hemocyte, such as bacteria. If phagocytosis does not clear the infection, nodulation occurs as the next response. This is a process whereby hemocytes bind to the pathogen and form aggregates of hemocytes and pathogens. The aggregate grows to form a small nodule within the hemocoel, within which pathogens are trapped and killed. Larger pathogens, such as parasitoids, are encapsulated within several layers of hemocytes. The humoral immune response is characterized by the production of antibacterial peptides; it is similar to the innate response of vertebrates. Hemocytes or the fat body produces these peptides, and their production continues for several weeks after the primary infection. This provides the insect with a degree of immunity against further attack. *See* CIRCULATION.

Muscles. Insect muscles are similar to those found in other organisms except that all insect muscles are striated; there are no smooth muscles. The skeletal muscles are connected to the exoskeleton by cuticle structures called apodemes. These muscles move the sclerites (hardened outer plates) of the insect, allowing the insect to move. Visceral muscles are found in most internal organs, including the gut, heart, and reproductive tissues. Skeletal muscles can be further differentiated into synchronous and asynchronous muscles. These differ based on motor neuron activity and contraction of the muscle. Synchronous muscles contract once for every neuronal impulse, whereas asynchronous muscles have several contractions per neural impulse. Synchronous muscles are found throughout the body and can be both skeletal and visceral. Asynchronous muscles are found only as flight muscles and occur in pairs. A feature of asynchronous muscles is that they are stretch-activated; thus one set of muscles contracts slightly, stretching the other pair. This stretching then activates the muscle for a contraction. In this manner the asynchronous muscles work together and can maintain a contraction-relaxation rate independent of neural input for several cycles. However, a neural input is required to maintain the contraction rate of asynchronous muscles over longer periods.

Asynchronous muscles are primarily used as flight muscles that are of the indirect kind, meaning that they do not directly move the wing; instead they are connected to the exoskeleton of the thorax. Contraction of the indirect muscles moves the body wall of the thorax, resulting in movement of the wing. There are two types of indirect flight muscles, dorsal-longitudinal and tergosternal muscles (**Fig. 2**). When the dorsal-longitudinal muscles contract, the top of the thorax moves up; and when the tergosternal muscles contract, the top of the thorax moves down. Thus these muscles operate alternately to cause movement of the wing. Direct muscles are attached to the wing in the hinge region and provide control of aerodynamic forces placed on the wing during flight. Insects that have asynchronous flight muscles also have the highest wing beat frequencies; for example, some small flies have wing beat frequencies greater than 500 beats per second. The indirect asynchronous muscles contract and relax once during each wing beat; thus there is the need for a discontinuous neural input. In contrast, insects that use synchronous flight muscles have wing beat frequencies typically lower than 50 beats per second.

A consequence of the high contraction rate of flight muscles is that they also have the highest metabolic demands of any muscles. High metabolic rates require the availability of oxygen, which is supplied by the tracheal system. In some cases the tracheoles actually penetrate deep within the muscle. The fuel to supply muscle contraction comes from adenosine triphosphate (ATP), but the amount of ATP present in resting muscles would last only 1 second during flight. Therefore, muscles must regenerate the ATP, and this takes place in the mitochondria. For most insects, sugar in the form of glycogen is broken down to provide energy for muscles. In those insects that are constantly taking in sugar, such as foraging honeybees, the diet supplies the necessary energy for the flight muscles. In those insects that undertake long-distance flight, such as migrating insects like the locust or monarch butterfly, stored lipid is utilized as a fuel source. The lipid is stored in the fat body as triacylglycerol and transported to the muscles through the hemolymph as diacylglycerol, which is carried by the protein lipophorin. This process is regulated by peptide hormones (adipokinetic hormones) produced in the corpora cardiaca near the brain.

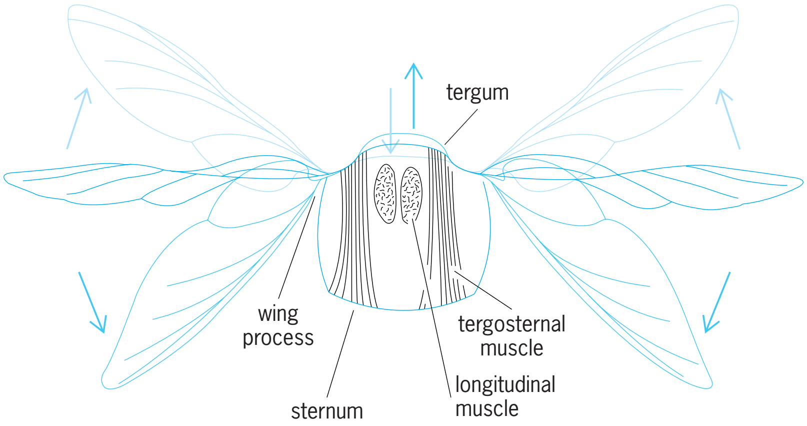


Fig. 2. Wing movements in the flight of an insect. (*After T. I. Storer and R. L. Usinger, General Zoology, 4th ed., McGraw-Hill, 1965*)

In some larger insects the thoracic flight muscles generate a considerable amount of heat, necessary to maintain the optimal temperature for muscle action. A preflight warm-up period is used to heat the muscles to this optimal temperature. In some large sphingid moths, this temperature is around 35°C (95°F), which can be considerably higher than the environmental temperature. During flight the thoracic temperature is kept from rising too high by a modification of the circulatory system in which excess heat is transferred to the hemolymph and pumped back to the abdomen. The abdomen is cooler, and the excess heat is lost to the environment. Thus, some insects can thermoregulate body temperature before and during flight. *See* FLIGHT; MUSCULAR SYSTEM.

Programmed cell death. Larval insects have intersegmental muscles that connect abdominal segments and are used in larval locomotion. During the ecdysis to the pupal stage, some of these muscles degenerate and die. After the adult emerges from the pupal case, the remaining intersegmental muscles die. The muscle cells that undergo programmed cell death (apoptosis) are regulated by the decline in the amount of ecdysone and 20-hydroxyecdysone. As in *Manduca sexta*, the decline in ecdysteroid amount results in the activation of certain genes and the repression of other genes in these muscles. Two of the repressed genes encode for actin and myosin, which are the contractile proteins of muscles. Two of the stimulated genes encode for apolipophorin and polyubiquitin. Polyubiquitin genes encode multiple forms of ubiquitin, which is a 76-amino-acid protein found in all eukaryotic cells. The main mechanism for the specific degradation of proteins is an ATP-requiring addition of ubiquitin to proteins. These tagged proteins are then rapidly degraded by the proteasome. Thus, cell death is not simply the cessation of muscle growth, but is due to the programmed development of specific pathways that lead to cell death. The loss of intersegmental muscles in *M. sexta* is a model system that has been used to study programmed cell death in eukaryotes. Other cells within the insect also undergo programmed cell death during metamorphosis, for example, specific neurons within the central nervous system.

Reproductive system. The internal structures of reproduction are the testis in the male and the ovaries in the female. The external reproductive structures are quite varied among the different types of insects and in some cases are the only structures that separate species.

Sperm and eggs. The internal male reproductive structures consist of the testis and tubes for carrying the sperm to the ejaculatory duct. In most cases there is also a seminal vesicle for storage of sperm. There are accessory organs that produce the seminal fluid that is transferred to the female with the sperm. In some insects the sperm and seminal fluid are encapsulated into a structure called a spermatophore. This spermatophore can be preformed before copulation or can be made during copulation. In some insects the spermatophore is used as a mating plug that blocks the genital tract of the mated female to help prevent other males from fertilizing the female.

The internal female reproductive structures consist of the paired ovaries connected to a common oviduct for laying eggs. Females also have accessory organs for producing material that is laid along with the eggs. A spermatheca is present for the storage of sperm. During mating, the spermatophore is first placed in a bursa copulatrix and then the sperm is transferred to the spermatheca. In some insects the sperm can be stored and remain viable in the spermatheca for the entire life of the female. This is especially so for the social queen female, which mates once and then fertilizes thousands to millions of eggs using sperm from the one male.

The process of egg development is called oogenesis. Each ovary is made up of a few to 100 ovarioles. In some insects each ovariole produces one egg at a time, while in other insects each ovariole has continual egg development. During egg maturation, yolk is formed inside the developing eggs. Most of this yolk is protein and lipid. The main protein that is found in the yolk is called vitellogenin. It is produced by the fat body, secreted into the hemolymph, and taken up by the ovaries through a process of selective endocytosis. This occurs before the chorion or eggshell is formed. The receptors on the ovarioles that aid in the uptake of vitellogenin have been identified at the molecular level, and they are similar to the low-density lipoprotein receptors found in mammals.

In general, there are two types of reproductive patterns of females. Females can continuously develop eggs, or they can produce eggs in discrete batches. In the latter group the synchronization of egg development is usually triggered by a meal or the laying of a previous batch of eggs. For example, some mosquitoes develop eggs in batches, and this step is governed by the taking of a blood meal. In most females, juvenile hormone is the main hormone involved in regulating egg development. Juvenile hormone stimulates the production of vitellogenin by the fat body and also the uptake of vitellogenin by the ovaries. In the Diptera (mosquitoes and flies), ecdysone is the hormone that stimulates the fat body to produce vitellogenin.

Fertilization. Once the egg has been fully developed, fertilization takes place just prior to being laid. The sperm enter the egg through the shell or chorion through special openings called micropyles. Oviposition, or the process of laying eggs, is a complex behavior that requires the right environmental conditions, and also the availability of a suitable substrate. For most insects, this involves a lot of visual, physical, and chemical cues to find the proper egg-laying substrate.

Mating also has a great influence on egg development and oviposition. The male in most cases is passing not only sperm to the female but also chemical information that will change her behavior. This behavior change usually results in increased oviposition and decreased receptivity toward other males. For example, *Drosophila* males pass a sex peptide to the female that changes her behavior and results in decreased sexual receptivity toward other males. Part of the decreased receptivity is a result of lower sex pheromone levels of the female. This helps ensure that the male's sperm will fertilize the eggs that the female lays. In crickets it has been shown that prostaglandins will stimulate oviposition in females. The male transfers both the substrate for prostaglandin formation and the enzyme prostaglandin synthetase to the female to stimulate an increased oviposition rate. In those insects where the female mates more than once, evolution has produced mechanisms whereby the male can increase the probability that his sperm will fertilize the eggs and his genes will be passed on to the offspring.

Most insects are oviparous, reproducing by laying eggs. A few insects are viviparous, reproducing by bearing live young. For example, in the tsetse fly, eggs hatch within the mother and the larvae receive nourishment from her until they are ready to pupate. The mother develops one larva at a time. The aphids are also viviparous and give birth to live young. Aphids are also parthenogenetic, meaning that reproduction can occur without fertilization. In the case of aphids, females are produced during the summer months parthenogenetically and then males are produced in the fall. Males and females mate and the females lay eggs that overwinter. The next year females hatch from the eggs to start the cycle over again.

Nervous system. Insects, like other higher organisms, have complex nervous systems. The nervous system processes information coming in from the environment via sensory systems mediating vision, touch, taste, smell, and hearing; and it produces a reaction to that information resulting in a typical behavior. The processing may result in motor neuron activity stimulating sets of muscles. For example, the tactile information that the feet are losing touch with the ground is sufficient to start the neuronal flight pattern. Wind-sensitive hairs on the tail

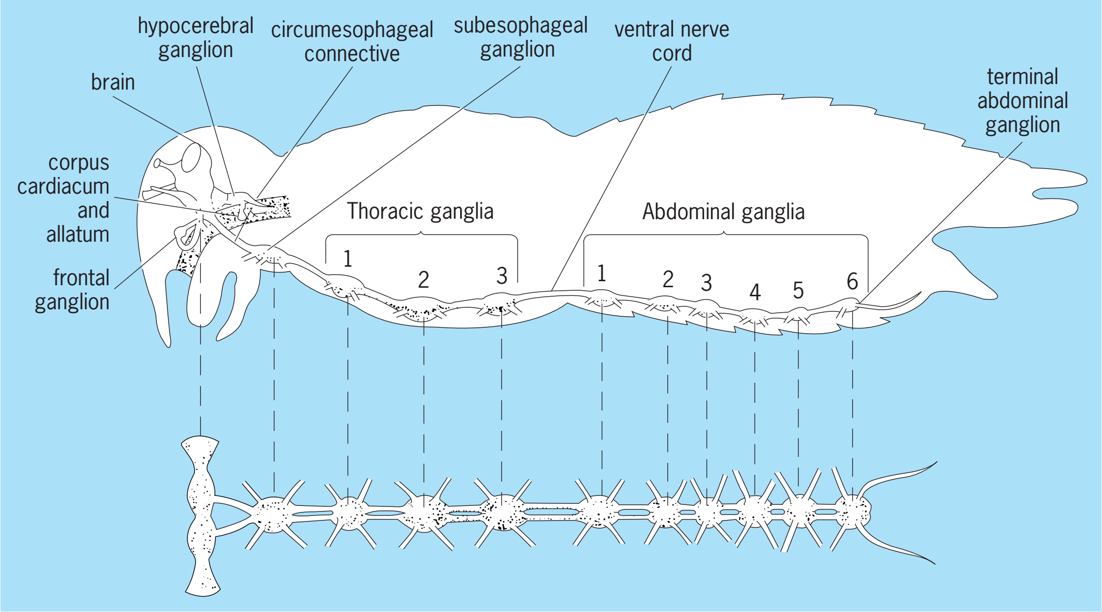


Fig. 3. Generalized insect nervous system. (*After W. S. Romoser and J. G. Stoffolano, Jr., The Science of Entomology, 4th ed., McGraw-Hill, 1998*)

end of a cockroach detect slight air movement and prompt the cockroach to run away. These events are integrated within the nervous system of the insect.

Morphologically the nervous system consists of segmental ganglia that are connected together by a paired solid nerve cord (**Fig. 3**). These ganglia are found ventral in the hemocoel, and in the primitive state there is one per segment. However, in most insects some of these ganglia have fused together. The brain is composed of several ganglia and is responsible for information coming from the eyes (vision), antennae (smell), and mouthparts (taste). The thorax usually has three ganglia, responsible for movements of the wings and legs. Each abdominal segment has smaller ganglia, which have control over the internal tissues and abdominal movement. These ganglia are all connected together by a paired ventral nerve cord. The ganglia themselves contain the individual nerve cells (neurons), with the axons traveling either through the ventral nerve cord or out to the periphery through nerves. Sensory neurons are located in the periphery and send their axons into the central nervous system. Sensory neurons are coupled together with motor neurons by interneurons. Together all of these neurons are involved in integrating information and regulating the behavior of the insect. *See* NERVOUS SYSTEM (INVERTEBRATE).

Vision. Adult insects usually have two kinds of light-detecting organs: compound eyes and ocelli (small simple eyes composed of photoreceptor cells and pigment cells). Hemimetabolous insects have compound eyes throughout development, whereas holometabolous insects have compound eyes as adults and stemmata during the larval stages. Stemmata are simple compound eyes that can discriminate shapes and detect movement. However, they probably are not involved in composing a visual image, and holometabolous larval insects rely more on chemical and mechanical receptors to learn about their environment.

The compound eyes of insects are composed of a few to thousands of individual light-gathering structures called ommatidia (**Fig. 4**). If only a few ommatidia are present, they are usually widely spaced and circular in pattern. If there are many ommatidia, they make up the hexagonal faceted eyes of most insects.

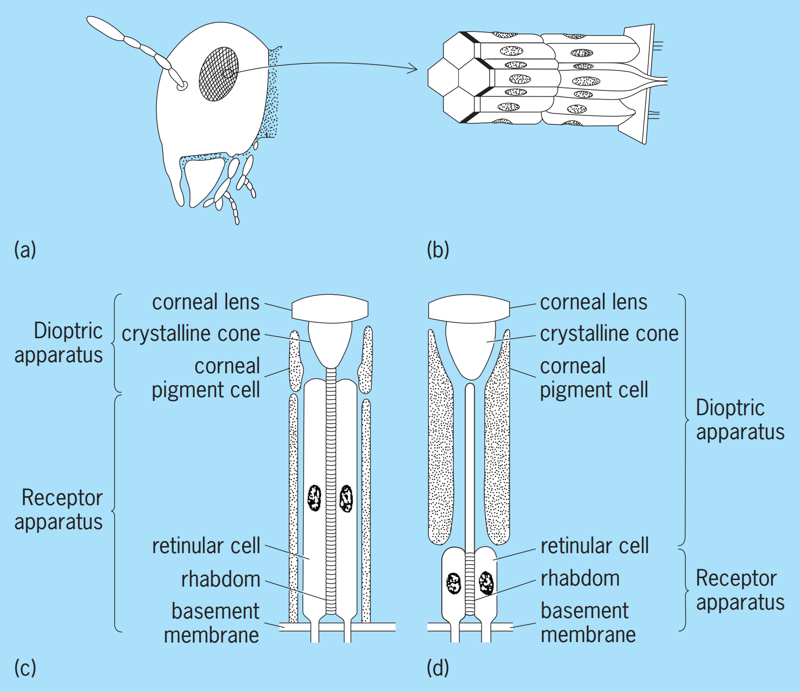


Fig. 4. Compound eye structure. (*a*) Head with compound eye. (*b*) Four ommatidia removed and magnified. (*c, d*) Representative types of ommatidia. (*After W. S. Romoser and J. G. Stoffolano, Jr., The Science of Entomology, 4th ed., McGraw-Hill, 1998*)

Light strikes photopigments in the ommatidia, and this information is transduced into a neural signal that is sent to the optic lobe of the brain. Neural centers within the brain process this information into an image that is thought to be right side up (superposition). Each ommatidium has a visual field of view of about 3–8°. However, the large number of ommatidia can cover a greater field of view, in some cases over 180°. Thus with both compound eyes an insect can detect images in a wide field of view without the need to turn the head.

The outer portion of an individual ommatidium has a biconvex corneal lens that is actually cuticle and an underlying crystalline cone (Fig. 4). These structures focus the light onto sensory cells called retinular cells. The corneal pigment cells are primarily for isolation of light between individual ommatidia to prevent light from bouncing around below the cornea. These pigments can move up or down to allow more light to be detected in low-light conditions such as at night. There are usually eight retinula cells arranged radially with light-detecting portions located in a central structure called a rhabdom. The retinular cells are monopolar sensory neurons that send their axons into the optic lobe of the brain. Light-sensing pigments known as rhodopsins are located in the rhabdom and can detect different wavelengths of light. Most insects detect blue, green, and ultraviolet wavelengths, with some detecting red wavelengths.

Some insects that fly at night, such as the moths, have developed a clear-zone type of eye, or a superposition eye (Fig. 4*d*). The pigment in the corneal pigment cells moves up around the crystalline cone, leaving a clear area between the cone and the retinular cells. This happens in the dark-adapted eye to allow more light to be detected. Light entering one ommatidial lens can cross over and be detected by adjacent ommatidia. Moths also have a tapetum, in which tracheae run parallel to the retinular cell and reflect light back into it. This allows more light to be detected but is thought to decrease the resolution of the final image. However, the image resolution is apparently still good enough to allow the insect to fly at night.

The other type of eye is an apposition eye, and it is characterized by the retinular cells located just below the crystalline core (Fig. 4*c*). The corneal pigment cells help to concentrate light onto the rhabdom of individual ommatidia. Apposition eyes are typically found in insects that fly during the day (diurnal).

Visual acuity is hard to judge in insects, but in general it is thought to be inferior to that of humans. However, insects such as the honeybee can discriminate certain shapes and patterns. Most insects can detect polarized light coming from the Sun using the ultraviolet light receptors and the orientation of the rhabdoms in the retinular cells. Detecting polarized light is important in helping the insect determine direction. Insects also have the ability to detect a high frequency of light flashes. This is important for fast-flying insects to detect objects and for stationary insects to detect other insects flying.

Ocelli are visual organs found in most insects. They are usually paired on the dorsal surface of the head with one located medially. These have a single corneal lens with many retinular cells located beneath the lens. They do not form images but are used to detect the intensity of light. Flying insects use them to help determine the position of the horizon. *See* EYE (INVERTEBRATE).

Chemical sensilla and pheromones. The senses of taste and smell are involved in detecting chemicals in the environment. Smell is involved with detecting airborne chemicals (olfaction), and taste is involved with detecting chemicals in the food. Each has a unique type of sensory structure called sensilla. A sensillum is similar to a tactile hair where the cuticle forms a hair with an underlying sensory neuron and accessory cells. Contact sensilla involved in taste have one pore in the cuticle at the tip of the hair, whereas the sensilla detecting airborne chemicals have multiple pores in the cuticle hair. The shape of the hair can be varied: from hairlike structures found on antennae of moths, to short peglike structures on antennae of locusts, to pits on antennae of flies. Physiologically these sensory cells act in a similar manner. The compound being detected crosses the cuticle through a pore and comes in contact with a receptor lymph fluid. The receptor, which is a monopolar neuron, senses the compound, and this causes the neuron to depolarize and a signal is sent down the axon to the brain. It is unclear how the receptor neuron actually senses a specific compound, but it probably involves the compound binding to a receptor on the sensory neuron. Putative receptor proteins have been identified recently in insects using molecular cloning techniques. These proteins belong to the superfamily of receptors named G-protein-coupled receptors (GPCR), with characteristic seven-transmembrane domains. Identification of a wide variety of chemoreceptors has been aided by the recent sequence identification of several insect genomes.

Contact chemosensilla involved in the sense of taste can be found on various body parts: antennae, mouthparts, and feet. For example, the tarsi (feet) of flies have over 3000 contact chemosensilla for the detection of sugar in food. If these sensilla come in contact with a high enough concentration of sugar, the proboscis (tubular organ) extends and the insect probes with additional sensilla on the mouthparts. If sufficient stimulatory sensory information is present, the insect feeds. In addition to detecting stimulatory compounds, insects can detect deleterious compounds. For example, the fly readily feeds on a sugar solution, but if a certain concentration of salt is also present the fly does not feed. Most insects have similar types of sensilla for detecting a wide variety of compounds, and these compounds can stimulate or inhibit feeding. This is important for phytophagous (plant-feeding) insects that usually feed on only one type of plant. Plants have produced a wide variety of compounds to inhibit insect feeding, and in turn insects have developed ways of detecting these compounds. In some cases the insect eats the plant containing the deleterious compound anyway but must inactivate the compound in some manner.

Chemosensilla detecting airborne or volatile compounds are usually found on the antennae. These chemosensilla have hairlike cuticle structures with an underlying sensory neuron. The volatile compound, usually hydrophobic, enters the multiporous hair and comes in contact with the aqueous sensory lymph. In this sensory lymph are proteins that bind to the volatile compound. These odorant-binding proteins probably function in the transport of the volatile compound across the sensory lymph to the sensory neuron cell, where the volatile compounds are detected. Just like the contact chemosensory cell, the neuron then depolarizes and a signal is sent down the axon to the antennal lobe of the brain. The brain integrates this information and produces a behavioral response. The brain antennal lobe morphology is such that the integration areas are formed into ball-like structures called glomeruli. These neuronal glomeruli are groups of axons that originate from specific sensory hairs on the antennae. These sensory neurons relay the information to other neurons that send the information to higher-order processing centers within the brain. In moths a set of these glomeruli is found only in males and receives the information coming from specific sensory neurons on the antennae for detecting sex pheromones.

Some of the sensory neurons on the antennae are involved in detecting general types of odors such as host-plant odors. The other sensory neurons are involved in detecting specific types of chemicals. Pheromones are chemicals used by insects to communicate between individuals of the same species. In general there are four different types of pheromones: sex, aggregation, alarm, and trail. Sex pheromones are chemicals produced by one sex to attract an individual of the opposite sex. Aggregation pheromones are used to attract both sexes to a particular site for feeding or for mating. Alarm pheromones are released to signal other individuals of danger; these are primarily released by social insects and by aphids. Trail pheromones are released to mark a given area or territory and are usually less volatile than the other types of pheromones.

A considerable amount of research has been conducted on sex pheromones. Moths represent a group of insects that utilize sex pheromones, and in most cases it is the female that is producing the volatile sex pheromone to attract the male. Many factors are important in order for a male moth to find the female, including environmental factors such as time of day, wind speed, temperature, and humidity. The female produces a blend of chemicals that are usually species-specific, and the male moth uses this blend of chemicals to find the female. The female has a special pheromone gland that makes and releases the chemicals into the air. These chemicals float downwind, forming a plume of odors. The male detects these chemicals with specific chemosensilla on the antennae, which results in a typical upwind flight pattern to search for the female. Males are very sensitive to the sex pheromone molecules, and in some cases only a few molecules are detected. In the silkmoth it was calculated that only about 300 molecules need to be detected in order for the male to initiate the upwind flight behavior.

Most pheromones are a blend of components, most of which are needed to maintain a search flight for the female. There are also compounds, some of which are pheromones for other species, which inhibit male flight. An example is the gypsy moth and the nun moth. The gypsy moth female produces a chemical called (+)-disparlure (7*R*,8*S*-2-methyl-7,8-epoxyoctadecane). The nun moth female produces both (+)-disparlure and (−)-disparlure. These compounds are the same compound except the epoxide is an optical isomer (+ is 7*R*,8*S*; − is 7*S*,8*R*). The gypsy moth male can detect both isomers, with the (+)-isomer producing upwind flight and the (−)-isomer inhibiting upwind flight. In contrast, the nun moth male initiates upwind flight only if both isomers are present. This is a way to separate the two species so the male will find the right female, and sex pheromones help to ensure species separation.

There have been a variety of chemicals identified that insects utilize as pheromones. These can be quite complex to relatively simple structures. The first was identified in the silkmoth in 1959 as an alcohol with 16 carbons and two double bonds (*E*10,*Z*12-hexadecadien-1-ol). In some cases the chemical structure is of sufficient complexity that no other chemical in nature would have the same structure. Another strategy that insects utilize is to use a blend of chemicals that are specific to one species. The blend of chemicals becomes the pheromone. These chemicals can be volatile and airborne as occurs with the moths and beetles, or they can be nonvolatile contact chemicals as occurs with cockroaches and flies. In most cases the female produces the pheromone chemical in specialized glands, and the production of the pheromone can be under endocrine regulation. Juvenile hormone has been shown to regulate pheromone production in adult cockroaches and beetles. In adult flies it is either juvenile hormone or ecdysone. In the moths it is a peptide hormone produced in the subesophageal ganglion of the brain. *See* PHEROMONE.

Touch and hearing. The sense of touch in insects is concerned with detecting forces put on the body by mechanical energy. This energy may originate either internally through the action of muscles or externally by the environment. Perception of the various forms of mechanical energy results in the sensing of gravity, pressure, touch, position, vibration, and hearing. Sensory cells called mechanosensilla detect these mechanical forces. These sensilla are quite variable in structure, but they all possess a neuron with dendrites reaching the cuticle, and an axon that sends an impulse to the central nervous system. The variability is due to the number of neurons present and the type of cuticular structure. In addition, there can be both mechanosensilla and chemical sensilla in one structure. Essentially all the mechanosensilla can detect mechanical forces placed on the cuticle. When the cuticle is deformed, the dendrite senses this, resulting in depolarization of the membrane, and a neural impulse is sent down the axon to the central nervous system.

The cuticle is shaped differently for different types of mechanosensilla. The most common are the hair-shaped or trichoid sensilla. These are found all over the insect's body and are primarily involved as tactile receptors for touch. They also can detect air movement. In general a neural impulse is sent when the hair is bent and again when it is straightened. Hairs associated with joints have mechanosensilla where an impulse is sent continuously whenever the hair is bent. These are involved in determining positioning of the appendages. The head contains groups of these sensilla that are used to detect air movement. They are used to help determine wind direction when the insect is on the ground, and in flight to help control wing movement in relationship to the wind. Some of these hairs on the head are also involved in helping the insect detect the gravitational pull of the Earth. This is important in all insects to determine up and down, but especially important in the communication dance of honeybees, where up or down movement relays information to nest mates.

The campaniform sensilla are another type of mechanosensilla. These are characterized by the cuticle forming a cap to which the sensory neuron is attached. Again, these sensilla are detecting forces placed on the cuticle. These are usually found at the base of the wing and around other joints. Campaniform sensilla are important in helping the insect coordinate position of the legs while walking and wing position during flight.

A further type of mechanosensilla are the chordotonal organs that are responsive to vibrations. These sensilla may have no external sign of their presence. They are located beneath the cuticle but are attached to the cuticle by some kind of cell. If these sensilla are found positioned on the cuticle, in general they are involved in determining position of body parts. In some cases, these sensilla are tuned to detect certain sound frequencies produced by flying insects. For example, the larvae of the cabbage moth has eight sensilla on its thorax that can detect the vibrations produced by a flying yellow jacket wasp. When the larva hears that particular vibration, it stops moving or falls to the ground to avoid predation by the wasp. The other types of chordotonal organs are found in specific locations on the insect body. Some are found on the leg of crickets (subgenual organ) that function in hearing. These can detect vibrations coming from the substrate that the cricket is standing on. Others are found at the base of the antennae, forming the Johnston's organ. Insects use these to help determine position of the antennae, and in male mosquitoes they are used to detect the sound of flying female mosquitoes. Tympanal organs are the true organs of hearing in insects, and they are found on the thorax, legs, or abdomen. These organs have thin areas of cuticle to which chordotonal sensilla are attached. Sound waves vibrate the thin cuticle, and the vibration is picked up by the sensilla. These types of hearing organs are found in moths, grasshoppers, and cicadas. Moths use these organs to detect the echolocation cries of hunting bats, allowing the moths to take evasive action in flight. *See* PHONORECEPTION.

Sound and light production. Many insects communicate by producing sound. The sound can travel either through the air or through the substrate (the surface upon which an insect is standing). Many insects simply strike the substrate with a body part in a regular fashion to produce a vibration in the substrate. This is used as some kind of signaling mechanism to other insects of the same species. Another way of producing sound involves modifications of the cuticle surface. Stridulation refers to sound production by moving a cuticular scraper over a toothed file (a rib providing wing support, also called a vein). Crickets and grasshoppers produce their calling sounds in this way. Sound production in some crickets is produced by rubbing the scraper on one forewing along the file of the other forewing. Grasshoppers produce sound by rubbing the femur of the leg against the forewing. Another way to produce sound is exemplified by the high-pitched sound of the cicadas. These insects use a tymbal mechanism located in the first abdominal segment. The tymbal is a thin cuticle composed primarily of a flexible protein called resilin and a series of cuticular ridges. It is connected to a muscle that contracts to buckle the tymbal inward. This inward buckling produces the sound and, with repeated muscular contractions, yields sound patterns. Air sacs behind the tymbal amplify the sound and send it outward. Usually only the male produces sound to attract a conspecific (referring to individuals of a single species) female and for courtship behavior.

Several insects produce light without the aid of luminescent bacteria. Primarily several groups of beetles are known to produce their own light, notably the fireflies or lightning bugs. These beetles have a special organ usually found on the ventral side of the abdomen below a clear cuticle. The cells have a rich tracheal supply, providing the oxygen required for the light-producing reaction. A low-molecular-weight compound, known as luciferin, is the substrate for an enzyme called luciferase. This enzyme catalyzes a reaction using energy and oxygen to oxidize luciferin. The oxidized luciferin spontaneously decays and in the process gives off light. Millions of these reactions occurring simultaneously produce one flash. The regular flashes of the firefly are regulated by the nervous system and the amount of oxygen presented to the light organ cells. These regular flashes are used primarily in communicating to the opposite sex for mating purposes. *See* BIOLUMINESCENCE.

Russell Jurenka

Bibliography. R. F. Chapman, *The Insects: Structure and Function*, 1998; P. J. Gullan and P. S. Cranston, *The Insects: An Outline of Entomology*, 3d ed., 2005; M. J. Klowden, *Physiological*

Systems in Insects, 2002; J. L. Nation, *Insect Physiology and Biochemistry*, 2001; H. F. Nijhout, *Insect Hormones*, 1994; W. S. Romoser and J. G. Stoffolano, Jr., *The Science of Entomology*, 4th ed., 1998.

Insecta

A class of the phylum Arthropoda, sometimes called the Hexapoda. In fact, Hexapoda is a superclass consisting of both Insecta and the related class Parainsecta (containing the springtails and proturans). Class Insecta is the most diverse group of organisms, containing about 900,000 described species (see **table**), but there are possibly as many as 5 million to perhaps 20 million actual species of insects. Like other arthropods, they have an external, chitinous covering. Fossil insects dating as early as the Early Devonian have been found. *See* PARAINSECTA.

Approximate numbers of described species of insects in the modern world

Insect order	World species
Diplura	750
Archaeognatha	415
Zygentoma	320
Ephemeroptera	2,200
Odonata	5,100
Plecoptera	2,000
Grylloblattodea	25
Orthoptera	18,700
Phasmatodea	2,300
Mantodea	1,800
Blattodea	3,500
Isoptera	2,000
Dermaptera	1,500
Embioptera	2,000
Zoraptera	25
Psocoptera	2,600
Phthiraptera	3,500
Thysanoptera	8,800
Hemiptera	80,000
Megaloptera	260
Raphidioidea	100
Neuroptera	4,200
Coleoptera	340,000
Strepsiptera	370
Mecoptera	450
Siphonaptera	1,700
Diptera	102,200
Trichoptera	6,500
Lepidoptera	137,200
Hymenoptera	141,100

Classification

The class Insecta is divided into orders on the basis of the structure of the wings and the mouthparts, on the type of metamorphosis, and on various other characteristics. There are differences of opinion among entomologists as to the limits of some of the orders. The orders of insects (and their relatives the parainsects) are shown below.

Superclass Hexapoda
Class Parainsecta
Order: Protura; Proturans
Collembola; springtails
Class Insecta
Subclass Monocondylia
Order: Diplura; telsontails
Archaeognatha; bristletails
Subclass Dicondylia
Infraclass Apterygota
Order: Zygentoma; silverfish, firebrats
Infraclass Pterygota
Section Palaeoptera
Order: Ephemeroptera; mayflies
Odonata; damselflies and dragonflies
Section Neoptera
Hemimetabola
Order: Plecoptera; stoneflies
Grylloblattodea; rockcrawlers
Orthoptera; grasshoppers, katy-dids, crickets
Phasmatodea; walkingsticks
Mantodea; mantises
Blattodea; cockroaches
Isoptera; termites
Dermaptera; earwigs
Embioptera; webspinners
Zoraptera; zorapterans
Psocoptera; psocids, booklice
Phthiraptera; lice
Thysanoptera; thrips
Hemiptera; cicadas, hoppers, aphids, whiteflies, scales
Holometabola
Order: Megaloptera; dobsonflies, alderflies
Raphidioidea; snakeflies
Neuroptera; lacewings, antlions
Coleoptera; beetles
Strepsiptera; twisted-wing parasites
Mecoptera; scorpionflies
Siphonaptera; fleas
Diptera; true flies
Trichoptera; caddisflies
Lepidoptera; moths, butterflies
Hymenoptera; sawflies, wasps, ants, bees

Morphology

Insects are usually elongate and cylindrical in form, and are bilaterally symmetrical. The body is segmented, and the ringlike segments are grouped into three distinct regions: the head, thorax, and abdomen (**Fig. 1**). The head bears the eyes, antennae, and mouthparts; the thorax bears the legs and wings, when wings are present; the abdomen usually bears no locomotor appendages but often bears some appendages at its apex. Most of the appendages of an insect are segmented.

Body wall. In vertebrate animals the skeleton is on the inside of the body and is spoken of as the endoskeleton; in an insect the skeleton is primarily on the outside of the body and is called an exoskeleton. However, important endoskeletal structures occur, particularly in the head.

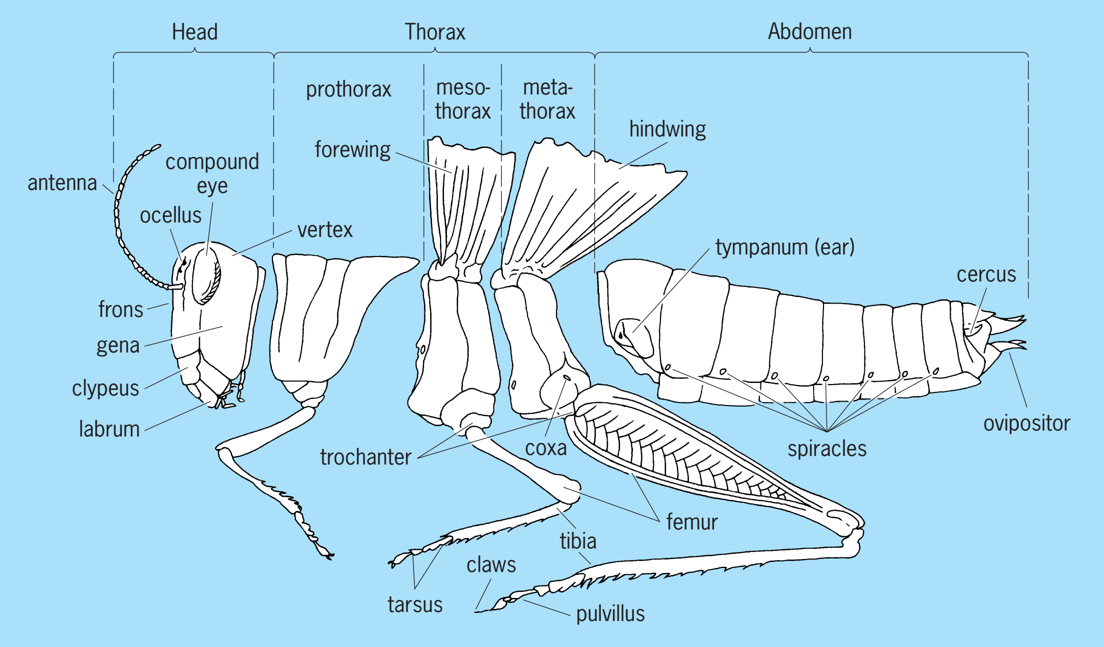


Fig. 1. Grasshopper, in lateral view, showing three main body divisions: head, thorax, and abdomen. (*After C. L. Metcalf, W. P. Flint, and R. L. Metcalf, Destructive and Useful Insects, 3d ed., McGraw-Hill, 1951*)

The body wall of an insect serves not only as a covering, but also as a supporting structure to which many important muscles are attached. The body wall of an insect is composed of three principal layers: the outer cuticula, which contains, among other chemicals, chitin; a cellular layer, the epidermis, which secretes the chitin; and a thin noncellular layer beneath the epidermis, the basement membrane. Chitin is a nitrogenous polysaccharide, $C_{32}H_{54}N_4O_{21}$, and is resistant to most weak acids and alkalies. It is secreted as a liquid and hardens. The surface of an insect's body consists of a number of hardened plates, or sclerites, separated by sutures or membranous areas, which permit bending or movement.

Head. On the head of an adult insect are the eyes, of which there are usually two types; the mouthparts, which are variously modified according to the feeding habits of the individual, and the associated digestive glands; and the antennae, which usually consist of several segments and vary greatly in size and form.

Eyes. A pair of compound eyes usually covers a large part of the head surface. The surface of each compound eye is divided into a number of circular or hexagonal areas called facets, each of which serves as a lens for a single eye unit, called an ommatidium. The compound eye may contain from a few hundred to several thousand facets. An eye of this type is believed to have mosaic vision, forming a great number of tiny similar images. In addition to having compound eyes, most insects also possess two or three simple eyes, the ocelli, usually located on the upper part of the head between the compound eyes; each of these has a single lens. The ocelli apparently do not serve as organs of vision, but rather are able to detect the degree of light intensity and its source. Insect larvae do not possess compound eyes, but usually have several ocelli. Adult insects never possess more than three ocelli. *See* EYE (INVERTEBRATE).

Mouthparts. Insect mouthparts typically consist of labrum, or upper lip; a pair each of mandibles and maxillae; a labium, or lower lip; and a tonguelike structure, the hypopharynx (**Fig. 2**). These structures are variously modified in different insect groups and are often used in classification and identification. The type of mouthparts an insect has determines how it feeds and what type of damage it is capable of causing. A cricket is said to have chewing mouthparts because it has heavily sclerotized mandibles that move sideways and can bite off and chew particles of food. A bedbug has long, slender, needlelike mandibles which can pierce the skin. Also, a pair of maxillae form a tube which is inserted through the puncture and used to suck blood.

Certain glands, such as the salivary glands, may secrete fluids which aid in digestion or which may serve as a defense mechanism against attack by other insects. These glands may also serve as a reservoir for disease organisms, such as malaria parasites. Modified salivary glands in silkworm larvae produce a fluid which, upon hardening, is known as silk.

Antennae. Several forms of antennae are recognized, to which various names are applied; they are used extensively in classification. The antennae are usually located between or below the compound eyes and are often reduced to a very small size. They are sensory in function and act as tactile organs, organs of smell, and in some cases organs of hearing.

Thorax. The thorax, or middle region, of the body bears the legs and the wings and is composed of three segments: the prothorax, mesothorax, and metathorax. Each thoracic segment usually bears a pair of legs; the wings, when present, are borne by the mesothorax and the metathorax. If only one pair of wings is present, they are usually borne by the mesothorax. The prothorax never bears wings, but in some fossil forms displays winglike structures. In some adult insects there are no wings, and in many immature and a few adult insects there are no legs. The thorax is usually connected to the head by a membranous neck region, the cervix.

Legs. The legs of modern insects typically consist of five general parts: the coxa, the basal segment next to the thorax; the trochanter, a small segment, occasionally two segments; the femur, usually a long segment and often quite large; the tibia, which may also be quite long; and the tarsus, the last portion, a series of small segments. The number of tarsal segments varies from one to five in different insects. The last tarsal segment usually bears a pair of claws and often a padlike structure between the claws. The legs may be variously modified and fitted especially for running, crawling, swimming, jumping, digging, or clinging to mammalian hair. The different segments of the leg may vary in size, shape, and spination. The crickets and long-horned grasshoppers have an eardrum, or tympanum, at the basal end of the front tibiae.

Wings. Insects are the only winged invertebrates, and their dominance as a group is probably due to their wings. Immature insects do not have fully developed wings, except in the mayflies, where wings develop in the subimago stage. The wings are saclike outgrowths of the body wall. In the adult they are solid structures and may be likened to the two sides of a cellophane bag that have been pressed tightly together. The form and rigidity of the wing are due to the stiff chitinous veins which support and strengthen the membranous portion. At the base are small sclerites which serve as muscle attachments and produce consequent wing movement. The wings vary in number, placement, size, shape, texture, and venation, and in the position at which they are held at rest. Adult insects may be wingless or may have one pair of wings on the mesothorax, or, more often two pairs, one each on the mesothorax and on the metathorax. The membranous wings may bear tiny hairs or scales. In some insects the front wings are thickened, leathery, or hardened and brittle. Most insects are able to fold the wings over the abdomen when at rest, but the butterflies, mayflies, dragonflies, and damselflies cannot do this; they hold the wings outstretched or together above the body when at rest. Some insects, such as the male crickets and grasshoppers, are able to produce characteristic sounds by rubbing the two front wings together or by rubbing the front wings with the hindlegs.

There is a common basic pattern of wing venation in insects which is variously modified and in general quite specific for different large groups of insects. Much of insect classification depends upon these variations. A knowledge of fossil insects depends largely upon the wings, because they are among the more readily fossilized parts of the insect body.

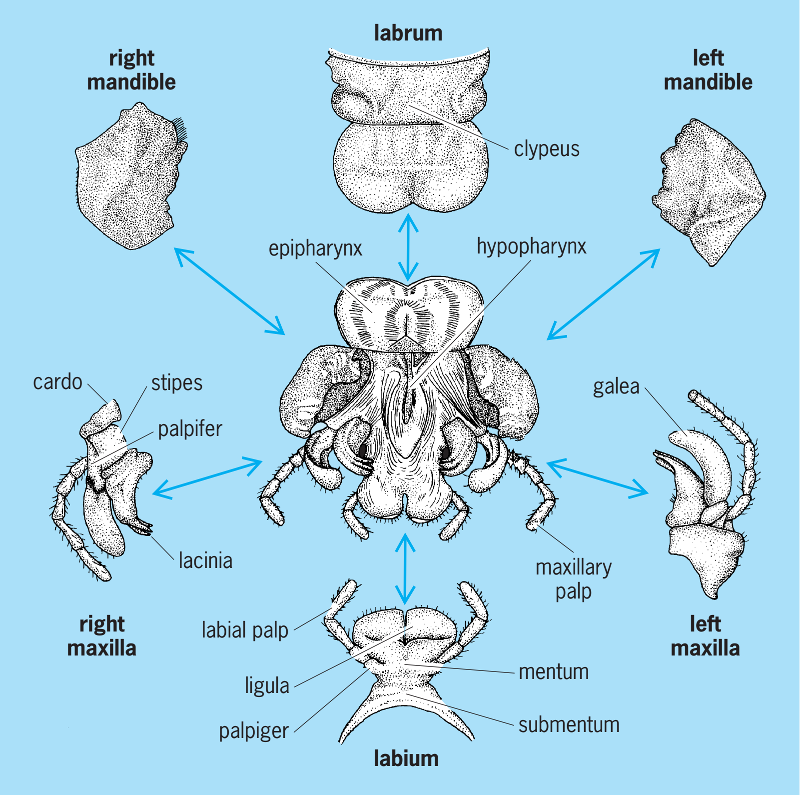


Fig. 2. **Grasshopper mandibulate mouthparts. (*After C. L. Metcalf, W. P. Flint, and R. L. Metcalf, Destructive and Useful Insects, 3d ed., McGraw-Hill, 1951*)**

Abdomen. The third division of the insect body is the abdomen; it has typically 11 segments. The eleventh segment is usually greatly reduced and is represented only by appendages; therefore, the maximum number of segments rarely appears to be more than 10. In many insects this number is reduced, either by a fusion of segments or by a telescoping of the terminal segments. Segments 1–7 bear appendages in various immature insects and in adult Archaeognatha, Zygentoma, and male Odonata. The abdominal appendages of immature insects may consist of tracheal gills (mayfly nymphs), lateral filaments (Neuroptera larvae), or prolegs (Lepidoptera larvae). These appendages are represented as styli in the Archaeognatha and Zygentoma and as copulatory structures in male Odonata. When 10 complete segments are present, a pair of clasperlike or feelerlike abdominal appendages, the cerci, arise from segment 10; however, they are absent in many species.

The genital structures arise from segments 8 and 9. In the male these function as copulatory structures for the transfer of sperm to the female. These external genital structures are quite variable and complex and frequently serve as excellent taxonomic characters. When not in use, these organs are usually withdrawn into the tip of the abdomen, often called the genital chamber, and frequently cannot be observed unless dissected. In the female these structures form the ovipositor or egg-deposition organ

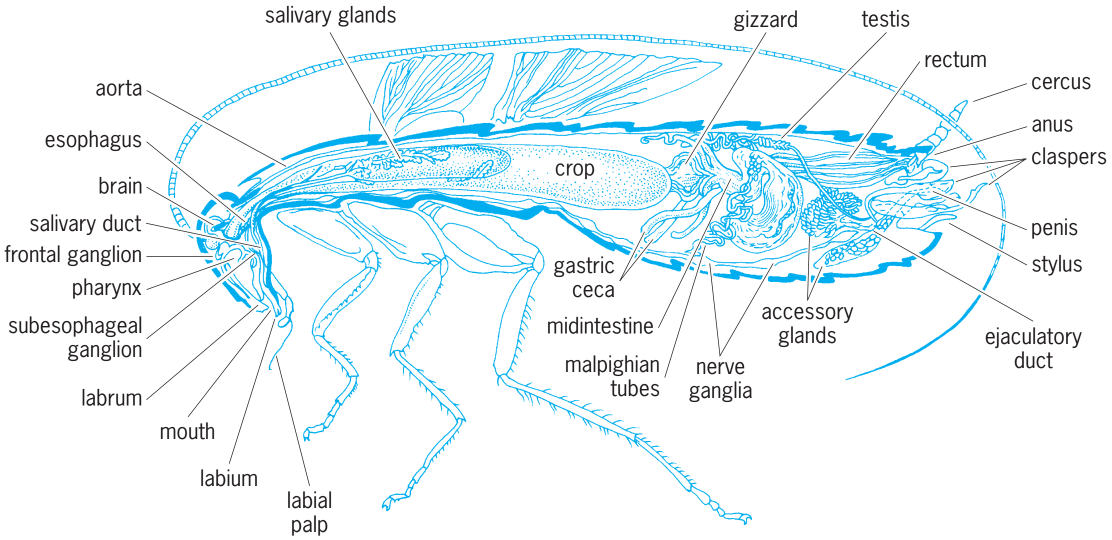


Fig. 3. Sagittal section of a cockroach showing the internal anatomy of a male specimen. (*After C. L. Metcalf, W. P. Flint, and R. L. Metcalf, Destructive and Useful Insects, 3d ed., McGraw-Hill, 1951*)

concerned with oviposition. Ovipositors are often fitted for slitting open plant or animal tissues and depositing eggs in the tissue. They may be modified to form a sting in the bees and wasps. In parasitic Hymenoptera they often are quite long and are used to drill into wood, where the eggs are deposited. In some of the meadow grasshoppers the ovipositor is a bladelike structure as long as the body.

Internal Anatomy

The internal structures of insects (**Fig. 3**) may be considered as six major systems: respiratory, digestive, excretory, circulatory, nervous, and reproductive.

Respiratory system. The intake of oxygen, its distribution to the tissues, and the removal of carbon dioxide are accomplished by means of an intricate system of tubes called the tracheal system. The principal tubes of this system, the tracheae, open externally at the spiracles. Internally they branch extensively, extend to all parts of the body, and terminate in simple cells, the tracheoles. The tracheae are lined with a thin layer of cuticle, which is thickened to form spiral rings that give the tracheae rigidity. The spiracles are located laterally and vary in number from 1 to 10. Usually there is a pair of spiracles on the mesothorax, a pair on the metathorax, and one pair on each of the first seven or eight abdominal segments. The oxygen enters through the spiracles, and carbon dioxide is eliminated principally through them; probably one-fourth of the carbon dioxide is eliminated through the body wall. In some insects the oxygen is taken in the body through the thoracic spiracles and is eliminated through the abdominal spiracles. The spiracles may be partially or completely closed for extended periods by some insects. Many adaptations for carrying on respiration are known. Aquatic insects may have tracheal gills located either on the external portion of the body or in the rectal cavity. The larvae of certain insects, such as mosquitoes, have respiratory tubes at the posterior end of the body. Some aquatic insects with normal tracheal systems and active spiracles can dive beneath the water and remain for extended periods of time by carrying with them a film or bubble of air on the surface of the body.

Digestive system. Insects possess an alimentary tract consisting of a tube, usually coiled, which extends from the mouth to the anus. It is differentiated into three main regions: the foregut, midgut, and hindgut. The foregut is usually composed of four parts: the pharynx, esophagus, crop, and proventriculus. The salivary glands are evaginations of the foregut. The foregut is an invagination of the ectoderm and is lined with cuticle. The midgut is usually undifferentiated, except for evaginations called gastric ceca, and is endodermal in origin and lined with epithelial cells. Some of the epithelial cells produce enzymes, and others absorb digested food. The hindgut may be differentiated into the small intestine, or ileum; the large intestine, or colon; and the rectum. The Malpighian tubules, which are excretory in function, connect with the alimentary canal at the point of union of the hindgut and midgut. Valves between the three main divisions of the alimentary canal regulate the passage of food from one region to another.

The major functions of the digestive system are the ingestion and digestion of food. Digestion is the process of changing food chemically so that it may be taken up by the blood to supply nutrients to various parts of the body. The food habits may vary in a given order. Larvae and adults of holometabolous insects usually have entirely different food habits and different types of digestive systems. Because of the great variety of insect foods, both the specific structures and the functions of the different parts vary greatly. When food is ingested, the crop may serve as a temporary storage chamber, where partial digestion may take place; before reaching the midgut, food passes through the proventriculus, which is often fitted with large or sharp needlelike teeth which may strain out or crush the larger particles of food. Digestion is completed and most of the absorption takes place in the midgut. After food is digested and for the most part absorbed, the residue passes into the hindgut, where some absorption may take place.

Excretory system. The excretory system consists of a group of tubes with closed distal ends, the Malpighian tubules, which arise as evaginations of the anterior end of the hindgut. They vary in number from 1 to over 100, and extend into the body cavity. Various waste products are taken up from the blood by these tubules and passed out via the hindgut and anus.

Circulatory system. The circulatory system of an insect is an open one, as compared with a closed system composed of arteries, veins, and capillaries, as in the vertebrates. The only blood vessel is a tube located dorsal to the alimentary tract and extending through the thorax and abdomen. The posterior portion of this tube, the heart, is divided into a series of chambers, each of which has a pair of lateral openings called ostia. The anterior part of the tube which lacks ostia is called the dorsal aorta. Pulsations of the heart produce the circulation by pumping the blood forward, and out of the aorta in the neck region. The increased pressure in this region then causes the blood to move posteriorly through the body cavity to the ostia.

Nervous system. The nervous system consists of a brain, often called the supraesophageal ganglion, located in the head above the esophagus; a subesophageal ganglion, connected to the brain by two commissures that extend around each side of the esophagus; and a ventral nerve cord, typically double, extending posteriorly through the thorax and abdomen from the subesophageal ganglion. In the nerve cords there are enlargements, called ganglia. Typically, there is a pair to each body segment, but this number may be reduced by the fusion of ganglia.

From each ganglion of the chain, nerves extend to each adjacent segment of the body, and also extend from the brain to part of the alimentary canal. There are several minor ganglia connected directly with the brain.

Reproductive system. Reproduction in insects is nearly always sexual, and the sexes are separate. Variations from the usual reproductive pattern occur occasionally. In many social insects, such as the ants and bees, certain females, the workers, may be unable to reproduce because their sex organs are undeveloped; in some insects, individuals occasionally occur that have characters of both sexes, called gynandromorphs. Also, parthenogenesis—the process of females giving rise to females—is known in some species.

The female reproductive system consists primarily of a pair of ovaries, each with a number of tubes called ovarioles. The ovarioles from each ovary unite into a common tube, and the two tubes fuse to form the oviduct. This discharges through an external opening, the vulva, located on the ventral side of the eighth segment. There are special glands and other chambers that open into the oviduct. The germ cell forms at the inner end of each ovariole and passes through a series of chambers, receiving its supply of food material as a consequence. In the last chamber it receives its chorion, or shell, which has a minute opening, the micropyle. As the completed egg proceeds into the oviduct, it passes the opening of a chamber, the spermatheca, where sperm previously received from the male is released to enter the egg and cause fertilization. The other glands may, at this point, contribute various substances used for attaching the egg to a substrate or for covering it.

In the male, there are paired testes in which the sperm are formed. From these, a pair of slender ducts, called the vasa deferentia, unite into a common ejaculatory duct, and this extends into a variously sclerotized penis. The sperm is commonly stored in swellings of the vasa deferentia to await copulation. Connected with these structures are various accessory glands.

Among the many groups of insects, various modifications exist in these basic structures and in their methods of operation. In fact, almost every method of reproduction that is known in the animal kingdom can be found among the insects. Copulation normally occurs through the vulva, but in the bedbug family Cimicidae of the order Hemiptera, it takes place by means of a special structure occurring at various points on the female body. The sperm is received in this structure and then passed through the body cavity to fertilize the egg in the ovariole. In some insects the sperm must pass through the oviduct into the ovariole to fertilize the egg before it enters the oviduct. The fertilized immature egg develops to maturity in the ovariole. In some forms the ovarioles are not tubular but are shaped like a bunch of grapes. In a few insects the embryo is retained within the body until it is fully grown, being nourished by glandular secretions. There is even something of a placental development in some forms of insect. *See* INSECT PHYSIOLOGY.

Metamorphosis

After insects hatch from an egg, they begin to increase in size (growth) and will also usually change, to some degree at least, in form and often in appearance. This developmental process is metamorphosis. The growth of an insect is accompanied by a series of molts, or ecdyses, in which the cuticle is shed and renewed. Prior to the actual shedding process, a new cuticular layer is secreted beneath the old layer, and a molting fluid is secreted which separates the new cuticle from the old.

The molt involves not only the external layers of the body wall, the cuticula, but also the cuticular linings of the tracheae, foregut, and hindgut; the cast skins often retain the shape of the insects from which they were shed. The shedding process begins with a splitting of the old cuticle, usually along the midline on the dorsal side of the thorax and head. This split grows and the insect eventually wriggles out of the old cuticle. The new skin, at first wrinkled and folded, remains soft and pliable long enough for the body to expand to its fullest capacity before hardening.

Insects differ regarding the number of molts during their growing period. Many have as few as four molts; a few species have 40 or more, and the latter continue to molt throughout life.

Insects have been grouped or classified upon the basis of the type of metamorphosis which they undergo. All entomologists, however, do not agree upon the same classification. These types of metamorphosis have been placed in three groups, under the headings ametaboly, hemimetaboly, and holometaboly. In order to cover the more detailed pattern, the following outline is used.

1. Ametabolous or primitive: No distinct external changes are evident with an increase in size.

2. Hemimetabolous: Direct metamorphosis that is simple and gradual; immature forms resemble the adults except in size, wings, and genitalia. Immatures are referred to as nymphs or naiads if aquatic.

3. Holometabolous: Complete, or indirect, metamorphosis; stages in this developmental type are: egg→larva→pupa→adult (or imago).

In the ametabolous forms, the insects never obtain wings and have no visible external changes. In the hemimetabolous forms, each insect gradually acquires wings during a period of growth and several moltings. In the holometabolous, or complete, form of metamorphosis, there is a complete change from the larva (which has only simple eyes, no wings, and frequently no legs or antennae) to the adult, usually of entirely different appearance, with wings, compound eyes, well-developed legs and antennae, and in certain groups entirely different mouthparts. Also, in this type of metamorphosis there is an extra life stage, the pupa, usually called a resting stage, in which the complete transformation takes place. The larva is the growth stage, the pupa the transformation stage, and the adult the reproductive stage. In the insects having complete metamorphosis, the feeding habits of the larva and the adult of the same species are usually different. Dwight M. De Long

Fossils

Insects and parainsects have a rich fossil record that extends to 415 million years, representing all taxonomic orders and 70% of all families that occur today. In addition, the insect fossil record documents an assemblage of extinct, higher-level taxa centered principally in the late Paleozoic. At a first approximation, the fossil insect record is divisible into a Paleozoic Insect Fauna of Early Devonian to late Permian age, and a Modern Insect Fauna beginning during the Triassic and continuing to their current unbridled diversity. These two evolutionary faunas are separated by the terminal Permian extinction of 251 million years ago, during which most species on Earth were extinguished, including many insect lineages. It also is likely that Permian taxa were being ecologically replaced by the progenitors of the Modern Insect Fauna, many of which survived the Permian extinction to continue as the familiar insect lineages of today (**Figs.** 4 and 5*a*).

Insect deposits are characterized by an abundance of exceptionally well-preserved deposits known as Lagerstätten. Lagerstätten refer not only to the familiar amber deposits that entomb insects in hardened tree resin (**Fig.** 6*a*), but more importantly to a broad variety of typically laminar, sedimentary deposits (Fig. 6*b*, *c*) that range from internally drained, fault-bounded montane lakes to lowland basins formed by broad tectonic subsidence. These lacustrine deposits, formed in lake basins, are the most persistent of insect-bearing deposits and document the evolution of insect biotas during the past 300 million years. By contrast, the oldest amber is approximately 120 million years old and extends modern lineages and associated taxa to the Early Cretaceous. Other major types of insect deposits include terrestrial shales and fine-grained sandstones marginal to marine deposits during the Early and Middle Devonian, a proliferation of nodular ironstone-bearing strata of late Carboniferous age from the equatorial lowlands of the paleocontinent Euramerica, and distinctive lithographic limestones worldwide from the Middle Jurassic to Early Cretaceous. More modern deposits are Miocene to Recent sinter deposits created by hydrothermal zones with mineral-rich waters, and similarly aged asphaltum, representing the surface accumulation of tar (Fig. 6*d*). Lastly, insects are abundant in many Pleistocene glacial deposits of outwash and stranded lake sediments, formed by the waxing and waning of alpine and continental glaciers.

Fossil record. Almost all the information of insect fossil history originates from approximately 150 major localities representing all the major types of deposits mentioned above. However, these deposits are distributed very unevenly in time and space. No insects are known from a 55-million-year hiatus between the Middle Devonian and the latest Mississippian. Another gap from the late Permian to Early Triassic features 15 million years of very poor, scrappy occurrences and an absence of any significant faunas. Described fossil insect deposits overwhelmingly occur at north temperate latitudes, principally North America and northern to central Eurasia, especially western Europe. Very few deposits have been found in the tropics, and only a few major deposits have been examined from south temperate latitudes, principally from southern Africa and eastern Australia.

Paleozoic. The earliest known insect is a bristletail (order Archaeognatha) that occurs in approximately 405-million-year-old Early Devonian deposits of Gaspé, Quebec; younger bristletails also are known from the Middle Devonian of New York state. An early Devonian collembolan from the 415-million-year-old Rhynie Chert of Scotland documents the closest relative of the insects, the Parainsecta. The Parainsecta (orders Collembola and Protura), when combined with the Insecta (all orders in Fig. 4), constitute a larger clade, Hexapoda, ancestors of which are thought to have originated from a marine, probably crustacean, arthropod sometime during the Late Silurian. During the Early Devonian, terrestrial ecosystems consisted of algae, mosses, primitive vascular plants, massive fungi, and arthropods that

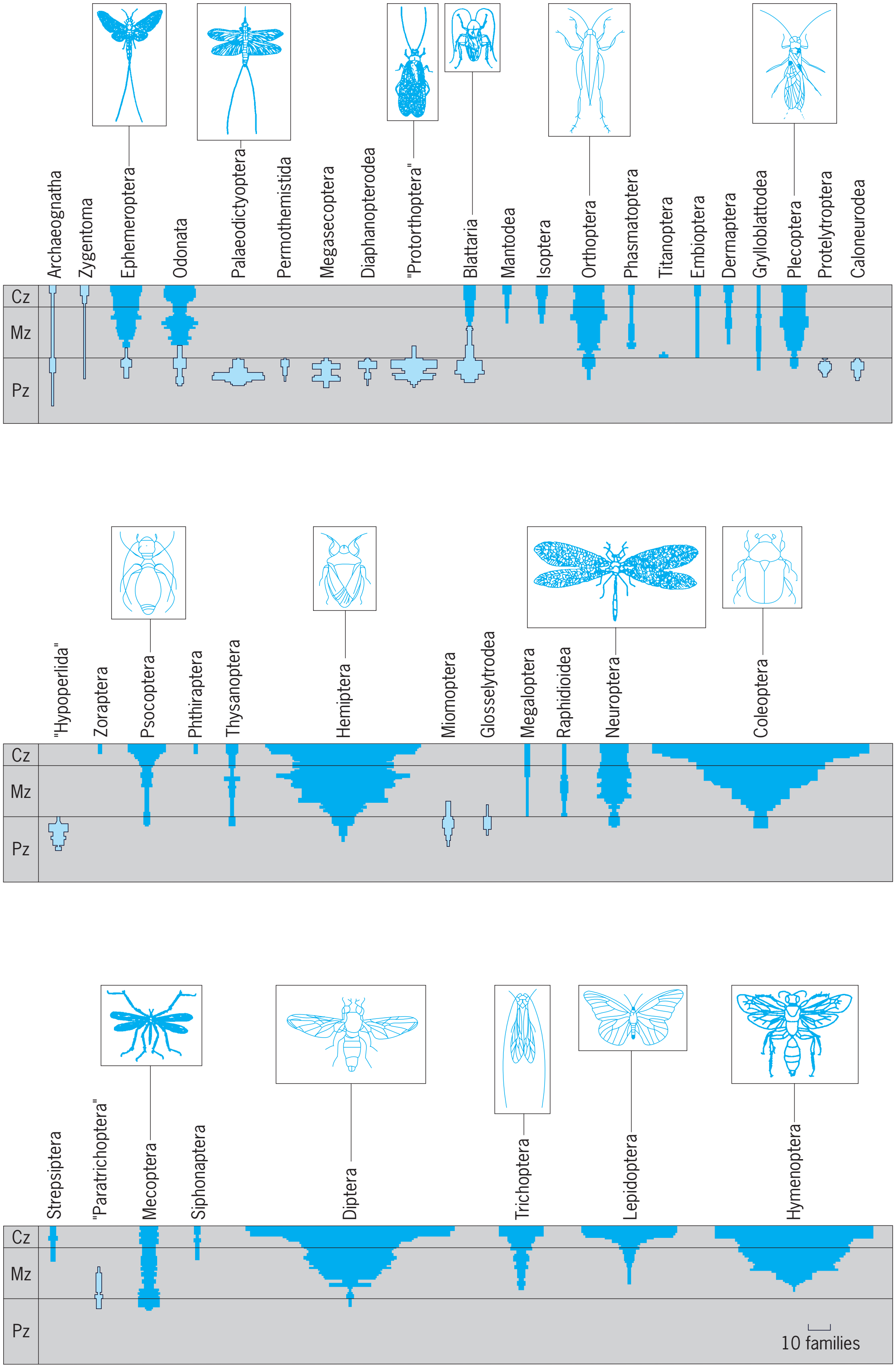


Fig. 4. Spindle diagrams showing diversities of fossil families within insect orders in stratigraphic stages of the Phanerozoic. Note the scale bar in the lower right. Abbreviations are Pz, Paleozoic (Silurian through Permian); Mz, Mesozoic; and Cz, Cenozoic. Boxed illustrations (not to scale) depict typical adult representatives of the more important orders; light-colored spindles or portions are Paleozoic Insect Fauna; dark-colored spindles or portions are Modern Insect Fauna. The appearance of angiosperms is approximately two-thirds of the way up for the Mesozoic band. (*Modified slightly from C. C. Labandeira and J. J. Sepkoski, Jr., Insect diversity in the fossil record, Science, 261:310–315, 1993*)

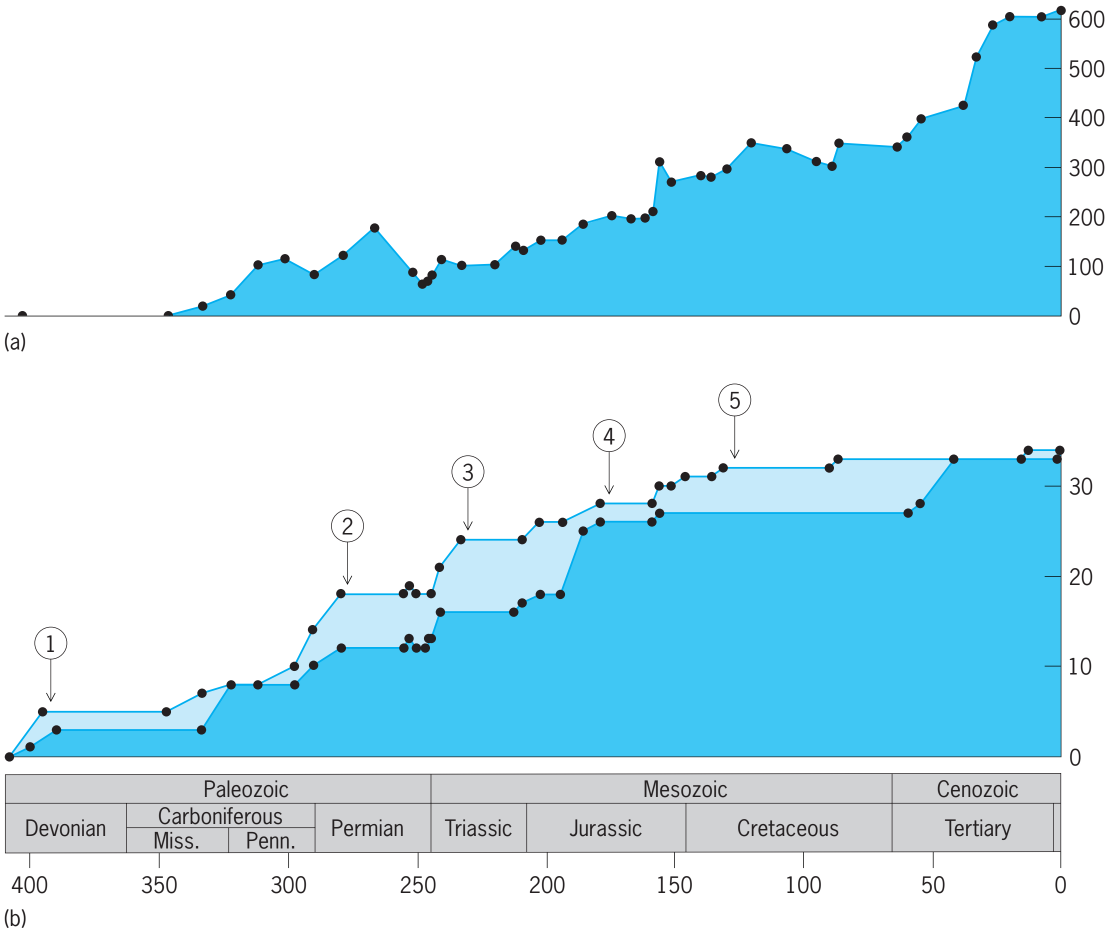


Fig. 5. Comparison of (*a*) family-level taxonomic diversity and (*b*) mouthpart class diversity since the Early Devonian. Data for both graphs are resolved at the geologic stage level, shown in *b* sequentially as data points but included in *a* only to reflect changes in diversity. For mouthpart diversity, in *b*, the darker pattern indicates strong evidence and the lighter pattern indicates less reliable evidence; numbers with arrows designate the five phases of mouthpart class diversification discussed in the text. Abbreviations are Miss., Mississippian; Penn., Pennsylvanian. (*Reproduced with permission from C. C. Labandeira, Insect mouthparts: Ascertaining the paleobiology of insect feeding strategies, Annu. Rev. Ecol. Sys., 28:153–193, 1997*)

included myriapods, arachnids, and primitively nonwinged collembolan and archaeognathan hexapods. Ecologically these early land habitats harbored photosynthesizing plants, detritivores, especially carnivores, but also significant herbivores, judging from all available evidence.

Permian insects are principally known from two deposits: Elmo, Kansas; and Chekarda, from the Ural Mountains of Russia. Many taxa from these localities occur in the same taxonomic families, suggesting that there were paleobiogeographical connections between the western equatorial and eastern north temperate margins of the vast, Northern Hemisphere continent of Laurussia. This Permian fauna represented a blend of scaled-down late Carboniferous lineages that were centered in the equatorial lowlands, and newly emergent, basal lineages of cicada-like hemipteroids and holometabolans occurring in the more arid higher latitudes. [Holometabolous insects (orders Miomoptera to Hymenoptera in Fig. 4) are characterized by immature larval stages separated from the adult stage by a typically resting stage, the pupa.] Hemipteroids (orders Psocoptera to Hemiptera in Fig. 4) and the holometabolans preferentially survived the end-Permian extinctions, forming the core of the Modern Insect Fauna during the Triassic.

After a 55-million-year hiatus, the known insect fossil record does not resume until the boundary between the early Carboniferous (Mississippian) and the late Carboniferous (Pennsylvanian), at 323 million years ago. At that time, numerous winged insect lineages appear in profusion and geochronologically suddenly (Figs. 4 and 5*a*), suggesting that there was either a significant but unrecorded earlier history, or a preceding spurt of rapid evolutionary diversification. By 300 million years, 15 orders and at least 100 families of insects were present, constituting the basis of the Paleozoic Insect Fauna. Most prominent by virtue of their large size (up to 76-cm wingspan) were dragonflies, mayflies, and four extinct palaeodictyopteroid orders with piercing-and-sucking mouthparts, known collectively as the Palaeoptera. With one exception, the Palaeoptera lacked the ability to swivel their laterally outstretched wings behind their abdomen. Other winged insects possess wing-folding mechanisms, including the second most abundant

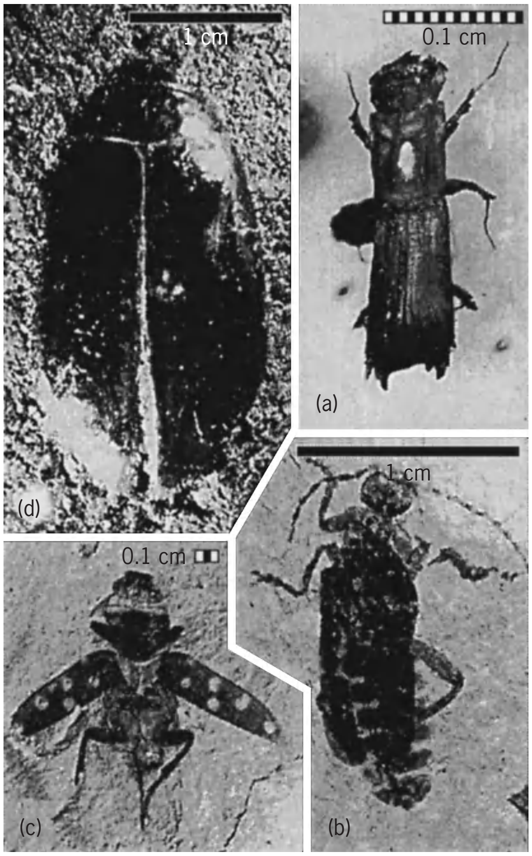


Fig. 6. Examples of the preservational variety in four well-known Lagerstätten deposits, illustrating Cenozoic beetles (Coleoptera). (*a*) Pinhole borer of the family Platypodidae, from early Miocene amber of the Dominican Republic. (*b*) Probable blister beetle of the family Meloidae, from the early Oligocene Antero Formation near Florissant, central Colorado. (*c*) Longhorn beetle of the family Cerambycidae, from the middle Eocene Green River Formation near Rifle, northwestern Colorado. (*d*) Predacious diving beetle of the family Dytiscidae, from the late Pleistocene asphalt deposit at McKittrick, south-central California. (*Paleobiology collections of the National Museum of Natural History, Washington, DC*)

Carboniferous group, the generalized and mandibulate Protorthoptera. The latter are an assemblage of many unrelated lineages that are difficult to define taxonomically, some of which have been segregated to the Hypoperlida and related Caloneurodea that combine a grasshopperlike body facies with mouthpart and head structures that resemble modern booklice (Psocoptera). Cockroaches were the third dominant constituent during the Carboniferous; although they were locally abundant, they had also a greater potential for preservation.

Mesozoic. Most insect clades present during the late Carboniferous did not survive the terminal Permian extinctions; of those that did, many eventually were extinguished during the early Mesozoic. Several clades present during the Permian, however, have survived to the present, either in decreased (Blattoidea, Mecoptera) or modestly expanded (Ephemeroptera, Odonata, Orthoptera, Thysanoptera, Neuroptera) diversity (Fig. 4). Notably, clades that originated during the late Permian to Middle Triassic, namely the Hemiptera and the three preeminent holometabolous orders with important fossil records (Coleoptera, Diptera, and Hymenoptera), overwhelmingly contributed to the Modern Insect Fauna, and their fossil diversity approximately matches proportionally their current global abundance. An exception is the Lepidoptera, whose earliest occurrence is Early Jurassic, and which possess a relatively poor fossil record when compared to their extant diversity.

This rebound is attributable to three major invasions of ecospace. (1) There was expansion during the Late Triassic to Late Jurassic of insect subadult stages (naiads and larvae) into fresh water as filterers, scrapers, shredders, and predators. This invasion provided a more complex trophic structure in aquatic ecosystems, reaching a modern type of organization during the Early Cretaceous. (2) On land, newly emergent lineages of seed plants, particularly advanced seed ferns, bennettitaleans, cycads, conifers, and ephedroids, were colonized by new holometabolous clades of phytophagous insects. These insect taxa consumed leaves, fructifications (fruiting bodies), and pollen, and included feeding types such as leaf miners, gallers, and wood borers that penetrated inner tissues. The establishment of virtually all of the broad categories of herbivory on seed plants during the Jurassic to Early Cretaceous was enlarged by the diversification and ecologic expansion of flowering plants during the mid-Cretaceous, which promoted at lower taxonomic levels further ecologic partitioning of food resources, and propelled new coevolutionary mutualisms. (3) The radiation of small wasp taxa during the Jurassic was associated with the emergence of new feeding guilds in which insects attacked other insects as food resources. This was done either by parasitoids eventually killing their prey, or by parasites feeding on body tissues without inducing death of the host.

Cenozoic. The terminal Cretaceous extinction, which extirpated planktonic foraminifera, ammonites, and dinosaurs, had no detectable effect on family-level insect diversity. However, like flowering plants, there now is evidence for a dieback of insect species, with rebound during the early Cenozoic. For herbivores, many modern-style associations with angiosperms are known from the early Cenozoic and even the Late Cretaceous. Additional evidence of prolonged longevity of lower-level insect taxa, when compared to other common invertebrates, includes the presence of extant genera in late Eocene (about 37 million years ago) Baltic amber, and the identification of species in early Miocene (about 21 million years ago) Dominican amber that are hardly distinguishable from modern species. Examination of fossil beetle species in Pleistocene glacial deposits reveals that there has been virtually no subsequent extinction, and that accommodation to major environmental perturbations has been by geographic dislocation rather than evolutionary adaptation.

Significance. Other types of fossil documentation are important for understanding insect paleobiology, such as the body-fossil history of mouthparts. A recent study of insect mouthparts reveals a fivefold phase of increasing mouthpart disparity through time (Fig. 5*b*). This geochronologic deployment of the 34 basic types of modern insect mouthparts began during the Early Devonian with a few generalized types (Phase 1), was expanded during the late Carboniferous to early Permian to include major modifications of mandibulate and piercing-and-sucking types (Phase 2), and increased significantly again during the Late Triassic to Early Jurassic (Phase 3) to include filter-feeding mouthpart types and others involved in the ecologic penetration of aquatic ecosystems, and also intricate interactions with other animal and seed-plant hosts. During the Late Jurassic, there was expansion of mouthpart types involved in fluid-feeding on plant, fungal, and animal tissues (Phase 4) and during the Early Cretaceous mouthpart innovation was completed by the addition of a few specialized mouthpart types involved in blood-feeding and other specialized associations (Phase 5). A comparison of taxonomic diversity and mouthpart disparity (Fig. 5) reveals that the generation of taxa has proceeded overall in a semilogarithmic increase reflected in a concave curve, whereas morphologic innovation, as revealed by mouthpart disparity, is a logistic process evidenced by a convex curve. This suggests that the deployment of basic morphologic types typically precedes taxonomic diversification in insect fossil history. Conrad C. Labandeira

Bibliography. S. A. Elias, *Quaternary Insects and Their Environments*, 1992; J. Kukalová-Peck, Fossil history and the evolution of hexapod structures, in I. D. Naumann et al. (eds.), *The Insects of Australia: A Textbook for Students and Research Workers*, 2d ed., vol. 1, pp. 141–179, 1991; C. C. Labandeira, Insects and other hexapods, in R. Singer and M. K. Diamond (eds.), *The Encyclopedia of Paleontology*, 1999; C. C. Labandeira, Plant-insect associations from the fossil record, *Geotimes*, 43:18–24, 1998; G. O. Poinar, Jr., *Life in Amber*, 1992.

Insecticide

A material used to kill insects and related animals by disruption of vital processes through chemical action. Insecticides may be inorganic or organic chemicals. The principal source is from chemical manufacturing, although a few are derived from plants.

Insecticides are classified according to type of action as stomach poisons, contact poisons, residual poisons, systemic poisons, fumigants, repellents, attractants, insect growth regulators, or pheromones. Many act in more than one way. Stomach poisons are applied to plants so that they will be ingested as insects chew the leaves. Contact poisons are applied in a manner to contact insects directly, and are used principally to control species which obtain food by piercing leaf surfaces and withdrawing liquids. Residual insecticides are applied to surfaces so that insects touching them will pick up lethal dosages. Systemic insecticides are applied to plants or animals and are absorbed and translocated to all parts of the organisms, so that insects feeding upon them will obtain lethal doses. Fumigants are applied as gases, or in a form which will vaporize to a gas, so that they can enter the insects' respiratory systems. Repellents prevent insects from closely approaching their hosts. Attractants induce insects to come to specific locations in preference to normal food sources. Insect growth regulators are generally considered to act through disruption of biochemical systems or processes associated with growth or development, such as control of metamorphosis by the juvenile hormones, regulation of molting by the steroid molting hormones, or regulation of enzymes responsible for synthesis or deposition of chitin. Pheromones are chemicals which are emitted by one sex, usually the female, for perception by the other, and function to enhance mate location and identification; pheromones are generally highly species-specific.

Inorganic insecticides. Prior to 1945, large volumes of lead arsenate, calcium arsenate, paris green (copper acetoarsenite), sodium fluoride, and cryolite (sodium fluoaluminate) were used. The potency of arsenicals is a direct function of the percentage of metallic arsenic contained. Lead arsenate was first used in 1892 and proved effective as a stomach poison against many chewing insects. Calcium arsenate was widely used for the control of cotton pests. Paris green was one of the first stomach poisons and had its greatest utility against the Colorado potato beetle. The amount of available water-soluble arsenic governs the utility of arsenates on growing plants, because this fraction will cause foliage burn. Lead arsenate is safest in this respect, calcium arsenate is intermediate, and paris green is the most harmful. Care must be exercised in the application of these materials to food and feed crops because they are poisonous to humans and animals as well as to insects.

Sodium fluoride has been used to control chewing lice on animals and poultry, but its principal application has been for the control of household insects, especially roaches. It cannot be used on plants because of its extreme phytotoxicity. Cryolite has found some utility in the control of the Mexican bean beetle and flea beetles on vegetable crops because of its low water solubility and lack of phytotoxicity.

The use of inorganic insecticides has declined to almost nil. Domestic production of inorganic arsenicals has apparently ceased, with current uses being supplied from existing stocks and by some limited importation. Probably the largest current uses are as constituents of formulations for wood impregnation and preservation.

Synthetic organic insecticides. After 1945, with the introduction of DDT, organic insecticides began to supplant the arsenicals. There are three major classes: the organochlorines, the organophosphorus insecticides, and the carbamates.

Organochlorines. The first of the chlorinated hydrocarbons to be widely used was DDT, followed by lindane. Since 1968, however, in the United States all nongovernmental and nonprescription uses of DDT have been canceled. Coincidentally, the use of lindane has declined both through the development of resistant strains of insects and through official restriction. Over the years, other chlorinated hydrocarbon insecticides have become available, although a number of these exhibited undesirable characteristics. They include TDE, methoxychlor, Dilan, chlordane, heptachlor, telodrin, aldrin, dieldrin, endrin, toxaphene, endosulfone, Kepone, mirex, chlordecone, and Perthane. Registrations for some have been canceled.

The U. S. Environmental Protection Agency (EPA) monitors the risks to human health and the environment of available insecticides and issues registrations specifying permitted uses. For example, registrations for all uses of TDE were canceled in 1971 because it is a metabolite of DDT. Most uses of aldrin and dieldrin were canceled by the EPA in 1975, and their legal use is limited to control of subterranean termites. Kepone and mirex were prohibited for all uses in 1976. Most uses of chlordane and heptachlor were canceled in 1978. Toxicity reviews for insecticides are scheduled as needed.

Organophosphorus insecticides. The development of this type of insecticide paralleled that of the chlorinated hydrocarbons. Since 1947, many thousands of organophosphorus compounds have been synthesized in academic and industrial laboratories throughout the world for evaluation as potential insecticides. Parathion and methyl parathion are two examples that have achieved significant use.

A great diversity of activity against insects is found among organophosphorus insecticides. It has been estimated that more than 100,000 different organophosphorus insecticides have been synthesized and evaluated. In general, they are very reactive species and thus readily degrade, having low soil persistence. Unfortunately, many can be quite poisonous to humans and other warm-blooded animals, as well as to insects. A few materials possess the desirable attribute of low mammalian toxicity, however. The gamut of available commercial products includes azinphosmethyl, bromophos, carbophenothion, chlorpyrifos, crotoxyphos, diazinon, dichlorvos, EPN, ethion, ethoprop, fensulfothion, fonofos, malathion, methidathion, naled, phosmet, sulfotep, stirofos, sulprofos, temephos, terbufos, and trichlorfon.

Schradan was unique among organic insecticides in that it showed systemic properties when applied to plants. Its activity upon direct contact with insects was relatively low, but following spray application to plants, it was absorbed, transported, and altered chemically by enzymatic processes to one or more products which were toxic to pests, such as aphids, which suck juices from plants. Schradan was thus a selective insecticide because it killed pests which fed on fluids from treated plants but did not affect predators which lived on the surfaces and preyed only upon the pests; for example, aphids are the target pests, and ladybird beetles and their larvae which prey upon the aphids are not affected by the insecticide retained within the plant. Unfortunately, schradan was ultimately found to have environmental properties, the risks of which exceeded benefits to the extent that in 1976 all registrations for use in the United States were canceled.

The following materials which exhibit a varied spectrum in degree and breadth of activity have become available commercially for insect control through systemic action following application to plants: acephate, demeton, dimeton methyl, dicrotophos, dimefox, dimethoate, disulfoton, methamidophos, mevinphos, monocrotophos, oxydemeton-methyl, oxymeton methyl, phorate, phosalone, and phosphamidon.

Coumaphos was one of the first products to be discovered which possessed systemic properties. In addition to coumaphos and dimethoate, chemicals useful as animal systemic insecticides include: chlorofenvinphos, chlorothion, crufomate, dialiflor, diaxothion dicapthon, diphos, famphur, fenthion, fenithrothion, menazon, methamiprophos, and ronnel.

Activity of organic phosphate insecticides results from the inhibition of the enzyme cholinesterase, which performs a vital function in the transmission of impulses in the nervous system. Inhibition of some phenyl esterases occurs also. Inhibition results from direct coupling of phosphate with enzyme. Phosphorothionates are moderately active, but become exceedingly potent upon oxidation to phosphates.

Carbamates. These insecticides are relatives of the alkaloid physostigmine. Individual compounds tend to exhibit selective insecticidal activity. The first commercial carbamate insecticide, carbaryl was introduced during the mid-1950s and continues to enjoy broad use not only in the United States but worldwide. The more important products include aminocarb, bendiocarb, bufencarb, dimetilan, dioxacarb, formetanate, methiocarb, mexacarbate, promecarb, and propoxur.

Considerable success has been achieved in discovering carbamates which exhibit systemic action when applied to plants. The more important products with this property include aldicarb, carbofuran, methomyl, oxamyl, pirimicarb, and thiofanox.

Carbamates, like organophosphates, are cholinergic. Several not only interfere with cholinesterase, but may also inhibit one or more enzymes known as aliesterases. The binding between carbamate inhibitor and enzyme appears to be much more readily reversible than does that between enzyme and organophosphate inhibitor.

Other types of insecticides. During the mid-1960s, much interest developed in the possibilities of utilizing certain of their own secretions and hormones for control of insects. The secretions, named pheromones, are released by one sex, usually the female, and perceived by the other. All pheromones so far discovered are highly species-specific chemicals whose primary function is to facilitate mate finding and therefore maintenance and propagation of the species. Pheromones are volatile materials, and chemically are frequently unsaturated esters of low-molecular-weight acids in which the molecules tend to lie within a range of 8–20 carbon atoms. The number of unsaturations and the isomerization associated with each are largely responsible for both their specificity and the difficulty of chemical characterization. Attempts to use pheromones to control insects have so far met with limited success. A combination of pheromone and a delivery system has been registered for use against the pink bollworm in cotton. Its performance seems to indicate that the procedure offers considerable promise. *See* CHEMICAL ECOLOGY.

Hormones from insects have been investigated intensively and given careful consideration as potential insect control agents; examples are the molting hormone and the juvenile hormone. The molting hormone is a steroidal molecule which is generally considered too complex and costly for manufacture and use as a pesticide. The juvenile hormone, however, is a more simple aliphatic unsaturated triterpenoid methyl ester. Juvenile hormones II and III have been discovered and identified. The natural hormones have thus far not proved suitable for use in practical pest control. However, some promising synthetic derivatives of the juvenile hormones have been registered, including neotenin, R-20458, hydroprene, and methopreme. *See* ENDOCRINE SYSTEM (INVERTEBRATE).

Diflubenzuron, another chemical with a different mode of action, interferes with the action of chitin synthetase, an enzyme necessary for the synthesis of chitin, and as a consequence of that process, disrupts the orderly deposition of that vital element of the exoskeleton. It is registered for the control of the gypsy moth in hardwood forests of the eastern United States. A conditional registration was granted for the use, beginning during the 1979 season, of diflubenzuron for the control and possibly the eradication of the boll weevil from cotton in the United States.

Pyrethrum is one of the oldest insecticides. Because it is highly active and quite safe for humans and animals, extensive effort was made during the period 1910–1945 to determine its chemical structure. The analytical problem was complicated greatly because a number of active principles were ultimately found to be present. Those identified were designated pyrethrins I and II, cinerins I and II, and jasmolins I and II. Efforts began immediately after identification of these natural products to produce chemicals with modified structures through laboratory synthesis. These efforts culminated in the discovery of some of the most active insect control agents yet devised. The first product, allethrin, which was only partially synthetic, was introduced commercially during the early 1950s. Subsequent, entirely synthetic products have included dimethrin, fenothrin, resmethrin, and tetramethrin.

There are two analogs which are highly active with residues on plant surfaces showing significant degrees of persistence. Their structures differ considerably from those of the natural product. Both fenvalerate and permethrin have been granted conditional registrations in the United States for use in the control of pests of cotton.

Insecticides obtained from plants include nicotine, rotenone, pyrethrins, sabadilla, and ryanodine, some of which are the oldest known insecticides. Nicotine was used as a crude extract of tobacco as early as 1763. The alkaloid is obtained from the leaves and stems of *Nicotiana tabacum* and *N. rustica*. It has been used as a contact insecticide, fumigant, and stomach poison and is especially effective against aphids and other soft-bodied insects.

Rotenone is the most active of six related compounds found in a number of plants, including *Derris elliptica, D. malaccensis, Lonchocarpus utilis*, and *L. urucu*. *Derris* is a native of eastern Asia, and *Lonchocarpus* occurs in South America. The highest concentrations of active principles are found in the roots. Rotenone is active against a number of plant-feeding pests and has found its greatest utility where toxic residues are to be avoided. Rotenone is known also as derris or cubé.

The principal sources of pyrethrum are *Chrysanthemum cinerariaefolium* and *C. coccineum*. Pyrethrins, which are purified extracts prepared from flower petals, contain chemically different active ingredients. The pyrethrins find their greatest use in fly sprays, household insecticides, and grain protectants because they are the safest insecticidal materials available.

Synergists. These materials have little or no insecticidal activity, but increase the activity of chemicals with which they are mixed. Piperonyl butoxide, sulfoxide, MGK, propyl isomer, and sesamex are commercially available. These synergists have their greatest utility in mixtures with the pyrethrins. Some have been shown to enhance the activity of insecticides as well.

Insect resistance. The resistance of insects to DDT was first observed in 1947 in the housefly. By the end of 1967, 91 species of insects had been proved to be resistant to DDT, 135 to cyclodienes, 54 to organophosphates, and 20 to other types of insecticides, including the carbamates. By 1975, 203 species were known to be resistant to DDT, 225 to cyclodienes, 147 to organophosphates, 36 to carbamates, and 35 to other insecticides. Since numerous insects are resistant to more than one type of compound, the total number of species involved is 364 worldwide. Almost every country has reported the presence of resistant strains of the housefly.

During 1957 and 1958, many growers of cotton in the Southern states changed from toxaphene and γ-BHC to organic phosphorus chemicals because of the resistance of the cotton boll weevil to chlorinated hydrocarbon insecticides. By 1967, resistant strains of the cotton bollworm and the tobacco budworm had developed and proliferated to the extent that in numerous areas the use of chlorinated hydrocarbon insecticides was of doubtful value. During the same decade, larvae of three species of corn rootworms also developed resistance to three potent chlorinated insecticides—aldrin, dieldrin, and heptachlor—active against a wide spectrum of soil-inhabiting insects. The development of resistance to chlorinated hydrocarbon insecticides among several species of disease-transmitting mosquitoes not only continues to pose a threat to world health, but led a World Health Organization committee to report in 1976 that efforts to cope with vector-borne diseases and to eliminate malaria were being hampered most of all by problems arising from resistance to insecticides.

The need to replace the highly active chlorinated materials stimulated the development and use of organophosphorus and carbamate insecticides. Strains of several species developed which were resistant to widely used organophosphates. One of the notable examples was that of the cotton pest, *Heliothis virescens* (the tobacco budworm), and methyl parathion, for which laboratory tests showed a greater than 100-fold difference between susceptible and resistant strains. The production of cotton in parts of northeastern Mexico was practically abandoned and was seriously threatened in the lower Rio Grande Valley of Texas by the mid-1970s due to difficulties and costs of attempting to control this pest.

Certain of the synthetic pyrethroids have been shown under controlled laboratory conditions to induce the selection of resistant strains of mosquitoes. This knowledge has been obtained prior to their registration and release for large-scale commercial development. Perhaps with this forewarning, these products will be employed in the field in ways that will minimize selection for resistance and thus prolong their usefulness for insect control. The synthetic pyrethroids, like the natural pyrethrins, have a very high unit activity, but differ from the briefly lasting natural products in that they persist for moderate periods of time following application in the field. These two attributes—high activity and persistence—are important factors associated with the selection of resistant strains of insects, regardless of the chemical involved.

Formulation and application. Formulation of insecticides is extremely important in obtaining satisfactory control. Common formulations include dusts, water suspensions, emulsions, and solutions. Accessory agents, including dust carriers, solvents, emulsifiers, wetting and dispersing agents, stickers, deodorants or masking agents, synergists, and antioxidants, may be required to obtain a satisfactory product.

Insecticidal dusts and formulated for application as powders. Toxicant concentration is usually quite low. Water suspensions are usually prepared from wettable powders, which are formulated in a manner similar to dusts except that the insecticide is incorporated at a high concentration and wetting and dispersing agents are included. Emulsifiable concentrates are usually prepared by solution of the chemical in a satisfactory solvent to which an emulsifier is added. They are diluted with water prior to application. Granular formulations are an effective means of applying insecticides to the soil to control insects which feed on the subterranean parts of plants.

Proper timing of insecticide applications is important in obtaining satisfactory control. Dusts are more easily and rapidly applied than are sprays. However, results may be more erratic, and much greater attention must be paid to weather conditions than is required for sprays. Coverage of plants and insects is generally less satisfactory with dusts than with sprays. It is best to make dust applications early in the day while the plants are covered with dew, so that greater amounts of dust will adhere. If prevailing winds are too strong, a considerable proportion of dust will be lost. Spray operations will usually require the use of heavier equipment, however. Whatever the technique used, the application of insecticides should be correlated with the occurrence of the most susceptible or accessible stage in the life cycle of the pest involved. By and large, treatments should be made only when economic damage by a pest appears to be imminent.

Attention has focused sharply on the impact of the highly active synthetic insecticides upon the total environment—humans, domestic animals and fowl, soil-inhabiting microflora and microfauna, and all forms of aquatic life. Effects of these materials upon populations of beneficial insects, particularly parasites and predators of the economic species, have been critically assessed.

The study of insect control by biological means has expanded. The concepts and practices of integrated pest control and pest management involve an insect control strategy that employs ecologically based procedures which rely on multiple interventions with minimal disturbances to the ecosystem.

Among problems associated with insect control are the development of strains of insects resistant to insecticides; the assessment of the significance of small, widely distributed insecticide residues in and upon the environment; the development of better and more reliable methods for forecasting insect outbreaks; the evolvement of control programs integrating all methods—physical, physiological, chemical, biological, and cultural—for which practicality was demonstrated; the development of equipment and procedures to detect chemicals much below the part-per-million and microgram levels. As a consequence of the provisions of the Federal Insecticide, Fungicide, and Rodenticide Act as amended by the Federal Environmental Pesticide Control Act of 1972, there have been increased efforts to obtain data delineating mammalian toxicology, persistence in the environment, and immediate chronic impact of pesticides upon nontarget invertebrate and vertebrate organisms occupying aquatic, terrestrial, and arboreal segments of the environment.

The registration of pesticides is a detailed, highly technical process. Pesticides must be selected and applied with care. Recommendations as to the product and method of choice for control of any pest problem—weed, insect, or varmint—are best obtained from county or state agricultural extension specialists. Recommendations for pest control and pesticide use can be obtained from each state agricultural experiment station office. In addition, it is necessary to read carefully and to follow explicitly the directions, restrictions, and cautions for use which are on the label attached to the product container. Insecticides are a boon to the production of food, feed, and fiber, and their use must not be abused in the home, garden, farm field, forest, or stream. *See* FUMIGANT; INSECT CONTROL, BIOLOGICAL; INSECT PHYSIOLOGY; INSECTA; PESTICIDE.

George F. Ludvik

Bibliography. A. J. Burn et al. (eds.), *Integrated Pest Management*, 1988; H. J. Cottrell (ed.), *Pesticides on Plant Surfaces*, 1987; P. Hurst, N. Dudley, and A. Hay, *The Pesticide Handbook*, 1989; D. H. Hutson and T. R. Roberts (eds.), *Insecticides*, vol. 5, 1986; *Kirk-Othmer Encyclopedia of Chemical Technology*, 4th ed., vol. 13, 1994; A. R. Leslie, *Integrated Pest Management*, 1994; R. L. Metcalf and R. A. Metcalf, *Destructive and Useful Insects: Their Habits and Control*, 5th ed., 1992; F. H. Nijhout, *Insect Hormones*, 1994, reprint 1998; H. H. Shorey and J. T. McKelvey (eds.), *Chemical Control of Insect Behavior*, 1977; C. Worthing and R. Hance, *Pesticide Manual*, 1991.

Insectivora

An order of placental mammals including shrews, moles, and hedgehogs. The tree shrews (Tupaiidae) and elephant shrews (Macroscelididae) are now recognized as unrelated, and they are placed in separate orders (Scandentia and Macroscelidea). Formerly thought to be the basal placental order, from which other orders were derived, the Insectivora is now restricted to members of the former suborder Lipotyphla. It evolved side by side with the other placental orders, with a fossil record going back to the Paleocene. A number of fossil families from the Cretaceous and early Tertiary, formerly included in the Insectivora, are classified as Proteutheria. *See* MACROSCELIDEA.

Description. Living lipotyphlous insectivores are small animals: the largest (*Potamogale*) weighs about 1 kg (2 lb). Most eat insects, worms, and other invertebrates, for which they search in ground litter and vegetation, using their highly developed olfactory sense and their mobile, sensitive snouts. Some burrow, such as moles; some are aquatic, such as the desman. Anatomically, they are distinguished by the absense of a cecum on the intestine, reduction of the pubic symphysis (fibrocartilaginous union of bones), and characters of the skull. The cheek teeth typically have sharp cusps and crests, and the incisors are often enlarged to act as forceps. Insectivores are found on all continents except Australia and Antarctica, but only one genus (*Cryptotis*, a shrew) has reached South America.

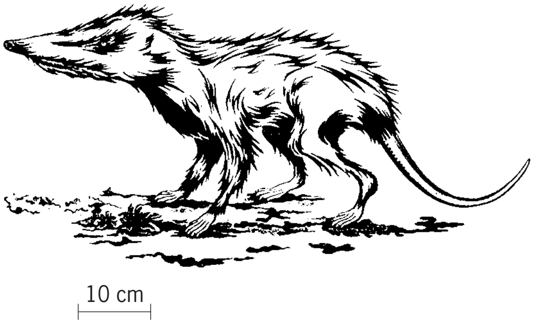


Fig. 1. Reconstruction of the giant erinaceid *Deinogalerix*, from the Miocene of Italy. (*From M. Freudenthal, Scripta Geologica, 14:1–19, 1972*)

Classification. Three suborders can be distinguished: Erinaceomorpha, Soricomorpha, and Chrysochloromorpha. Living erinaceomorphs belong to the family Erinaceidae, comprising the spiny hedgehogs (Erinaceinae) of Eurasia and Africa and the hairy moonrats (Echinosoricinae) of Southeast Asia. Probably originating in Asia, both subfamilies appeared in Europe in the Oligocene and in North America and Africa in the Miocene. *Proterix* (late Oligocene) and *Brachyerix* (late Miocene–Pliocene) are aberrant North American erinaceids which represent other invasions. Erinaceids died out in North America in the Pliocene. A giant echinosoricine, *Deinogalerix* (**Fig. 1**), with a skull 21 cm (8 in.) long, evolved in the late Miocene in isolation on a Mediterranean island, now part of Italy. A variety of erinaceomorphs existed in the early Tertiary of North America, Europe, and Asia, but most are known only by teeth and jaws. Exceptionally well preserved material of *Macrocranion* from the middle Eocene of Messel, Germany, shows the presence of spines on the back.

Four living families are included in the Soricomorpha: Soricidae (shrews), Talpidae (moles, desman), Tenrecidae (Madagascan tenrecs and African otter shrews), and Solenodontidae (*Solenodon*, confined to Cuba and Hispaniola). The phyletic relationships between these families remain to be elucidated. Dentitions of primitive Paleocene soricomorphs from North America and Europe resemble those of primitive erinaceomorphs. Soricids and talpids were of northern, probably Asiatic origin, and they were distinct before the end of the Eocene. Most early soricids belong to the subfamily Heterosoricinae, which became extinct in the Miocene. Shrews did not enter Africa until the middle Miocene. Among the talpids, the subfamily Talpinae is the most advanced in modification of the forelimbs for burrowing, the Desmaninae have aquatic adaptations, and the Uropsilinae (represented by the extant Chinese shrew mole) retained shrewlike limbs. True moles did not reach North America until the Miocene; they were preceded in the Oligocene by the Proscalopidae, which burrowed in a different way, with the hands moving longitudinally beneath the body, instead of laterally as in true moles. The Dimylidae, from the Oligocene and Miocene of Europe, are poorly known forms with crushing dentition; they may be talpid relatives.

The living *Solenodon*, in the West Indies, is one of the largest insectivores, but nevertheless shrewlike. In the Pleistocene it was accompanied by *Nesophontes* (Nesophontidae), which may be related.

They are believed to have become isolated in the West Indies since the early Tertiary. Attempts to relate them to mainland North American Eocene and Oligocene soricomorphs such as *Apternodus* and *Centetodon* have been unsuccessful. *See* SOLENODON.

The Tenrecidae are far removed geographically from other soricomorphs. They probably evolved in Africa, where their fossil history goes back to the early Miocene, and the otter shrews (Potamogalinae) survive on the continent of Africa today. The remainder are in Madagascar where, like the lemurs, they have evolved in diverse directions during a long time of isolation. In some ways, for example brain size, they have remained more primitive than other insectivores.

The golden moles (Chrysochloridae) are put into a separate suborder, Chrysochloromorpha. They are highly specialized burrowers, using large claws on the forefoot. They are confined to Africa, where fossils show that they were already specialized in the early Miocene.

Proteutherians. A number of more or less primitive placentals, from the Cretaceous and early Tertiary, that could not be incorporated in the better-defined orders were formerly classified in the Insectivora. With restriction of the Insectivora to the Lipotyphla, the order Proteutheria was proposed to include these fossils. This is an artificial and temporary expedient; as the phyletic relations of early placentals become better known, the proteutherians are being separated into more clearly defined orders.

The placentals had a long history in Asia prior to the extinction of the dinosaurs, going back into the Early Cretaceous of Mongolia (*Prokennalestes*). The best-preserved specimens are from the Late Cretaceous of the Gobi Desert. These include the primitive *Kennalestes* (**Fig. 2**) and *Asioryctes*, and the Zalambdalestidae, which were more specialized. The latter have been grouped on dental grounds with some lesser-known genera from central Asia in a Cretaceous suborder, Mixotheridia. Other central Asiatic fossils have been interpreted as early condylarths.

The Palaeoryctidae appeared in North America near the end of the Cretaceous and survived to the Eocene; they have been thought to be ancestral to creodonts, and some similarities to soricomorph insectivores have been noted. The Leptictidae (Late Cretaceous–Oligocene, mainly in North America) were insectivorous, ground-living animals, which have been claimed as distant relatives of lipotyphlan insectivores. The Pseudorhynchocyonidae, from the European Eocene, had long hindlegs, like kangaroo rats; they are possibly related to leptictids. The Apatemyidae (Paleocene–Oligocene) had enlarged incisors, probably used to obtain wood-boring insects, as in the living aye-aye; they are often put in a separate order, Apatotheria. The Pantolestidae (Paleocene–Oligocene), aquatic forms that reached a relatively large size and may have fed on mollusks, have also been separated as the order Pantolesta. The Paleocene Mixodectidae have been linked with the Dermoptera. *See* MAMMALIA; MOLE (ZOOLOGY); SHREW; TENREC. Percy M. Butler; Malcolm C. McKenna

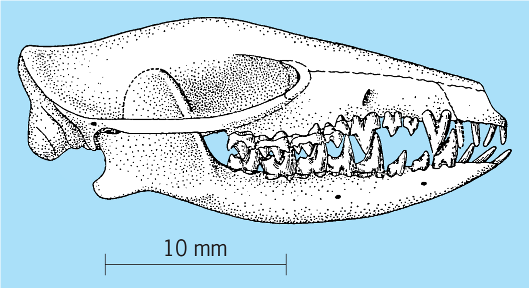


Fig. 2. Skull of *Kennalestes*, a proteutherian from the Cretaceous of Mongolia.

Bibliography. P. M. Butler, Phylogeny of the insectivores, in M. J. Benton (ed.), *The Phylogeny and Classification of the Tetrapods*, vol. 2, pp. 117–141, Clarendon Press, Oxford, 1988; G. C. Gould, Hedgehog phylogeny (Mammalia, Erinaceidae)—the reciprocal illumination of the quick and the dead, *Amer. Mus. Novit.*, 3131:1–45, 1995; S. B. McDowell, The Greater Antillean insectivores, *Bull. Amer. Mus. Nat. Hist.*, 115:113–214, 1958; R. D. E. MacPhee and M. J. Novacek, Definition and relationships of Lipotyphla, in F. S. Szalay, M. J. Novacek, and M. C. McKenna (eds.), *Mammal Phylogeny*, Part 2: *Placentals*, pp. 13–31, Springer-Verlag, New York, 1992; R. D. E. MacPhee, M. J. Novacek, and G. Storch, Basicranial morphology of Early Tertiary erinaceomorphs, and the origin of primates, *Amer. Mus. Novit.*, 2921:1–42, 1988; R. M. Nowack, *Walker's Mammals of the World*, Johns Hopkins University Press, 1999; C. A. Reed, Locomotion and appendicular anatomy in three soricoid insectivores, *Amer. Midland Natural.*, 45:513–671, 1951.

Insectivorous plants

Plants having variously modified, highly specialized leaves which capture and digest insects. The proteins of the digested insect bodies supply nitrogen, which otherwise may be unavailable to these plants in the places where they grow. Sometimes they are also called carnivorous (flesh-eating) plants. *See* NEPENTHALES; PITCHER PLANT; SECRETORY STRUCTURES (PLANT); SUNDEW; VENUS' FLYTRAP. W. W. Watson; Perry D. Strausbaugh

Insolation

The incident radiant energy emitted by the Sun, which reaches a unit horizontal area of the Earth's surface. The term is a contraction of incoming solar radiation. About 99.9% of the Sun's energy is in the spectral range of 0.15–4.0 micrometers. About 95% of this energy is in the range of 0.3–2.4 μm; 1.2% is below 0.3 μm, and 3.6% is above 2.4 μm. The bulk of the insolation (99%) is in the spectral region of 0.25–4.0 μm; about 40% is found in the visible region of 0.4–0.7 μm and only 10% is in wavelengths shorter than the visible. Energy of wavelengths shorter than 0.29 μm is absorbed high in the atmosphere by nitrogen, oxygen, and ozone.

Insolation depends on several factors: (1) the solar constant—that is, the amount of energy that in a unit time reaches a unit plane surface perpendicular to the Sun's rays outside the Earth's atmosphere, when the Earth is at its mean distance from the Sun; (2) the Sun's elevation in the sky; (3) the amount of solar radiation returned to space at the Earth–atmosphere boundary; and (4) the amount of solar radiation absorbed by the atmosphere and the amount of solar radiation reflected at the lower boundary of the Earth. Insolation is commonly expressed in units of watts per square meter, or calories per square centimeter per minute, also known as langley/min. For instance, the mean value of the solar constant has been estimated as 1368 W/m^2 ($\sim$1.96 ly/min), and the average insolation in summer for a midlatitude clear region could be 340 W/m^2 (700 ly/day), while for a cloudy region it is only about 120 W/m^2 (250 ly/day).

In the past, the solar constant was measured from balloons, rockets, aircraft, or high-altitude observations. Continuous extraterrestrial measurements of the solar constant are mainly from the instruments on the *Nimbus 6* and 7 satellites launched in mid-1975 and late 1978, respectively. These measurements gave the first clear evidence of short-term solar irradiance changes of the order of a few tenths of a percent. Longer-term variability is difficult to establish, since the longest continuous record from *Nimbus* 7 is about 10 years.

Since 1980, long-term solar monitoring has been done with the Active Cavity Radiometer Irradiance Monitor (ACRIM) on board the *Solar Maximum Mission* satellite. It is believed that the most accurate solar constant monitoring is from the self-calibrating ACRIM instrument. The ACRIM data have shown that the Sun's luminosity is negatively correlated with sunspots, which have an 11-year cycle. This is in contradiction to what was believed previously. Since the solar irradiance supplies the energy to drive the atmospheric and oceanic motions, it regulates the Earth's mean temperature and climate. A 1% change in the solar constant was found by some modelers to lead to a different climate. *See* CLIMATE HISTORY; METEOROLOGICAL SATELLITES; SOLAR CONSTANT.

The Sun's elevation, the complement of the zenith distance, depends on latitude, solar declination, and local hour angle. Therefore, the irradiance at the top of the atmosphere will have for each time of the year a certain diurnal variation and a meridional distribution which, in the annual mean, decreases toward the poles from a maximum at the Equator.

The Earth–atmosphere system reflects to space a certain fraction of the incident solar radiation, referred to as the planetary albedo. Even clear sky reflects to space by Rayleigh scattering from molecules, by reflection from the surface, and by scattering from aerosols (Mie scattering). The scattered radiation is called diffuse or sky radiation, and the radiation arriving directly from the Sun is called extraterrestrial direct or beam radiation. However, clouds affect the reflected portion of the insolation more than any other atmospheric factor. The cloud albedo depends on the drop size, liquid water content, water vapor content and thickness of the cloud, and the Sun's elevation. The smaller the drops and the greater the liquid water content, the greater the cloud albedo. The albedo of clouds varies from less than 10% to above 90%. From satellite observations it was determined that, when averaged over the globe, about 32% of the insolation is returned to space at the top of the atmosphere. *See* ALBEDO.

Approximately 20% of the solar radiation is absorbed in the atmosphere by water vapor, carbon dioxide, ozone, oxygen, aerosols, and clouds. The spectral variability of the solar irradiance will affect the amount and altitude of solar radiation absorbed by the Earth's atmosphere and can impact upper-atmospheric dynamical and photochemical processes. *See* ABSORPTION OF ELECTROMAGNETIC RADIATION; UPPER-ATMOSPHERE DYNAMICS.

The fraction of insolation that is reflected upward at the ground (surface albedo) depends strongly on the solar zenith angle, the spectral distribution of the radiation reaching the ground, and the type of the underlying surface. For instance, the near-infrared albedo may be three times larger than the visible albedo, especially for vegetation. The standard technique is to measure the total albedo integrated over wavelength. Typical surface albedos range from 5% for oceans to 12% for forests and 30% for deserts; fresh snow can reflect up to 90% of the incident energy.

About 2000 stations around the world routinely measure insolation, and about 1000 of them report to the Radiation Data Center in Leningrad. In the United States an existing network of solar radiation–monitoring stations was upgraded in 1977. It is operated by the National Oceanic and Atmospheric Administration (NOAA) and consists of 38 stations. The measuring instrument is known as a pyranometer. To supplement ground observations of insolation, models have been developed to estimate it from ground-based or satellite-cloud observations. It is interesting to note that the gradient of absorbed radiation, which is the driving force of the circulation, is similar to the gradient of the extraterrestrial insolation. *See* ATMOSPHERE; SOLAR RADIATION; TERRESTRIAL RADIATION.

Rachel T. Pinker

Bibliography. M. Igbal, *An Introduction to Solar Radiation*, 1983.

Instanton

A solution to the (imaginary-time) nonlinear field equation which arises in Yang-Mills field theory—a modern nonlinear generalization of Maxwell electromagnetic theory that is believed to give a fundamental description of elementary particles and forces. The instanton solution carries information about quantum tunneling. (Initially, the instanton was called a pseudoparticle.) *See* GAUGE THEORY.

A dynamical model when analyzed within quantum mechanics gives rise to processes different from those seen in an analysis based on classical mechanics. Of course, since nature is quantum mechanical, only the former processes are truly physical, but the latter can be used to give an approximate, semiclassical description. Of specific interest here is the tunneling phenomenon whereby a quantum-mechanical system evolves along paths that are classically forbidden, in that the classical equation of motion does not possess solutions which follow the forbidden paths. Nevertheless, there does exist in the classical theory a spur of quantum-mechanically allowed transitions: if, in the classical equation of motion, time t is replaced by imaginary time ($t \rightarrow \sqrt{-1}t$), solutions which follow tunneling paths can be found. In other words, processes classically forbidden in real time become allowed in imaginary time. Moreover, the quantum-mechanical tunneling amplitude Γ is approximately given by the equation below, where

$$\Gamma \approx \exp\left(-\frac{1}{\hbar}I\right)$$

I is the (imaginary-time) action for the (imaginary-time) tunneling solution, and $\hbar$ is Planck's constant. Obviously only solutions of noninfinite action are relevant, since Γ vanishes if I diverges. Also, I must be positive, so that Γ disappears in the classical limit, $\hbar \rightarrow 0$. *See* ACTION; NONRELATIVISTIC QUANTUM THEORY; QUANTUM MECHANICS.

While the above ideas, as they relate to point particles, were widely appreciated from the early days of quantum mechanics under the names semiclassical and WKB theory, their relevance to fields, such as the Maxwell electromagnetic field and its generalization the Yang-Mills field, was not realized until the mid-1970s, when an imaginary-time, finite-action solution to the Yang-Mills nonlinear wave equation was found. First called a pseudoparticle, then renamed an instanton, this solution was quickly interpreted as evidence for quantum tunneling in the Yang-Mills theory. This tunneling in turn served to explain the mass difference between the pi and eta mesons, and predicted that nucleons (protons and neutrons) are not stable, although their decay rate is exponentially small, hence negligible. [The instanton (tunneling) mediated proton decay is independent from that predicted by theories that try to unify all fundamental interactions.] *See* FUNDAMENTAL INTERACTIONS; GRAND UNIFICATION THEORIES; QUANTUM FIELD THEORY; WENTZEL-KRAMERS-BRILLOUIN METHOD.

The instanton solution is also of interest to mathematicians, who independently had come to a study of Yang-Mills fields from a purely mathematical context called fiber-bundle theory, which combines differential geometry and topology. The original solution was generalized, and it was found that the (imaginary-time) Yang-Mills action of an arbitrary solution is an integral multiple of the action for the original solution. The integer is called instanton number by physicists and Chern-Pontryagin number by mathematicians. The reason for this quantization with classical field theory derives from the topological properties of Yang-Mills fields. The most general N-instanton solution has been found. *See* FIBER BUNDLE.

In addition to calling attention to unexpected physical processes, namely tunneling, the instanton solution has put into evidence the rich topological structure that is present in Yang-Mills theory. Contemporary research is concerned with understanding this structure, an activity in which both physicists and mathematicians are engaged in a rare collaborative effort. *See* ELEMENTARY PARTICLE; TOPOLOGY.

Roman Jackiw

Bibliography. D. S. Freed and K. K. Uhlenbeck, *Instantons and Four-Manifolds*, 2d ed., 1991; R. Rajamaran, *Solitons and Instantons*, 2d ed., 1989, paper 1987; M. A. Shifman, *Instantons in Gauge Theories*, 1994.

Instinctive behavior

A relatively complex response pattern which is usually present in one or both sexes of a given species. These responses have a genetic basis, are essentially unlearned, and are generally adaptive.

The qualifying terms are an important part of the definition, because each of the defining characteristics of instinctive behavior can be shown in certain circumstances and under certain conditions to be inappropriate. Because of this, some theorists have suggested that the terms instinct and instinctive behavior be stricken from the lexicon. However, instinctive behavior is still a useful term to describe intricate actions such as the complex behavior of the courting bower bird and the efficient and precise killing behavior of the predatory cat.

Instinctive behavior occurs when an animal has a particular internal state while it is in the presence of a specific external stimulation called a releaser or a sign stimulus. Neither the internal state nor the external stimulus alone is adequate for the elicitation of the response. Many animals show particular instinctive behaviors only during the mating season, when hormonal changes associated with sexual behavior sensitize specific portions of the central nervous system, which will then be active in the presence of the releaser. The external stimulus may be relatively simple or incredibly complex. In forms lower on the phylogenetic scale, the releasing stimulus may be deceptively simple. When the small fresh-water fish, the three-spined stickleback (*Gasterosteus*), is hormonally prepared for breeding, it attacks any other male stickleback that comes near its nesting area. The fish's attack is stimulated by the opponent's red belly. The fact that there is another male attached to that red stimulus is really irrelevant. Experimental work has shown that the nesting male will attack a very crude cardboard model of a fish as long as it has a belly area that is painted red. During the spring and summer, the European robin (*Erithacus rubecula*) threatens or attacks another that approaches the breeding area. Once again, the releasing or eliciting stimulus is red, a tuft of red feathers. The robin attacks a tuft of feathers though it stands alone on a post with no bird near. Some sign stimuli (releasers) can be designed so that they elicit an exaggerated response which may actually be maladaptive. The eggs of the oystercatcher (*Haematopus ostralegus*) are spotted and of modest size, befitting a modest-sized bird. These eggs elicit particular behaviors from the bird, including sitting on them and nestling over them. However, if the oystercatcher is presented with a many-spotted egg several times larger than its own eggs, the bird shows a preference for the larger egg and repeatedly attempts nesting behavior on it, while its own eggs are neglected. *See* REPRODUCTIVE BEHAVIOR.

The innate neural mechanisms which permit the animal to recognize stimuli as complex as a spotted egg, or even the redness of a belly, cardboard or real, must be complex indeed. The linking of the neural mechanisms for perception with those that control the rather complicated responses are, of course, far more complex. A number of instinctive behaviors have been subjected to physiological analysis. Predation in the cat may serve as an example. One obvious releasing stimulus for predatory behavior in the cat is a rat. If the neural system for predation is in an active state, the prey will elicit the typical attack response. It should be noted, however, that the apparently simple stimulus “rat” is, in fact, remarkably complex. It does not matter whether the rat is lying down, curled up in a ball, walking toward or away from the cat, or standing on its hindlegs and washing its whiskers. It also makes no difference if the rat is white, brown, or spotted black and brown; the cat reacts to the vast complexity of those stimuli with a predatory response. Further, the cat may respond to those complex stimulus patterns even though it has never seen or been near a rat during its entire life.

In the wild cat a combination of internal neural sensitivity and relevant external stimuli combine to produce neural action which results in the killing response. However, many pet house cats show no inclination to attack small animals and may live with a rat for a prolonged period. It would appear that prey catching is not in their behavioral repertoire. Yet it is possible to produce relatively precise prey catching by direct electrical stimulation of a number of points in that portion of the brain called the limbic system. The chase and kill occur to direct brain stimulation even though the experimental animal has been completely isolated from the time of its birth.

The prey-catching and killing response shows considerable consistency from one cat to another. The prey is chased until caught, usually with the paws, and is held down while the killing bite is made to the back of the neck. This relative behavioral constancy has led some investigators to refer to instinctive behavior as fixed action patterns. Such terms cannot be taken literally, however. It is true that the end result is the killing bite, but the actual behaviors shown between the sighting of the prey and the final kill may vary widely and are a function of the behavior of the prey. If the prey jumps up on something, the cat follows. The cat’s muscle action pattern is very different if it chases the prey to the right or to the left. If the prey runs under something, the cat will show flexible behavior by reaching under the obstacle and attempting to strike the animal with its paw. Certainly these behavior patterns are fixed only in the broadest sense.

Within limits, the instinctive behaviors can be modified by learning. There is evidence, for example, that some predators learn to attack their prey at the back of the neck because when held in that position the prey cannot counterattack. *See* MIGRATORY BEHAVIOR.

Kenneth E. Moyer

Bibliography. J. Alcock, *Animal Behavior*: *An Evolutionary Approach*, 6th rev. ed., 1997; N. R. Carlson, *Physiology of Behavior*, 7th ed., 2000; R. A. Hinde and J. S. Hinde, *Instinct and Intelligence*, 3d ed., 1987.

Instrument landing system (ILS)

A collection of discrete radio navigation aids used by pilots of all types of aircraft for approach guidance to a specific airport runway, especially during times of limited visibility. Typically, three components constitute the instrument landing system: a localizer, which provides lateral guidance; a glide slope, which gives vertical guidance; and one to three marker beacons, which give position fixes along the approach path (**Fig. 1**). A distance-measuring-equipment (DME) readout or a nondirectional, low-frequency radio beacon (NDB) is sometimes substituted for the outer marker beacon, usually located about 5 mi (8 km) from the runway. Details essential for use of the instrument landing system by the pilot are peculiar to the airport and runway. *See* DISTANCE-MEASURING EQUIPMENT.

The localizer, glide slope, and markers function by radiating continuous-wave, horizontally polarized, fixed beams of radio-frequency energy into the air space. The localizer operates in the band of 108–112 MHz, the glide slope in the band of 329–335 MHz, and the marker beacons on the discrete frequency of 75 MHz. The localizer and glide-slope frequencies are paired; for example, every localizer operating at 109.5 MHz has a glide slope operating at 329.6 MHz. If there is a DME present, it is also paired. Only the localizer receiver in the aircraft needs to be tuned.

Use. Over 1000 instrument landing systems operate in the United States, with an additional 300 systems in use worldwide. There are approximately 117,000 aircraft with ILS receivers, many with two for redundancy purposes. Engineering advances have made it possible to install the instrument landing system at all sites where the physical obstructions allow an instrument procedure to be developed. A major contender for ultimate replacement of the instrument landing system appears to be a Global Navigation Satellite System (GNSS), in which the Global Positioning System (GPS) would play a major role. (The microwave landing system may be used for

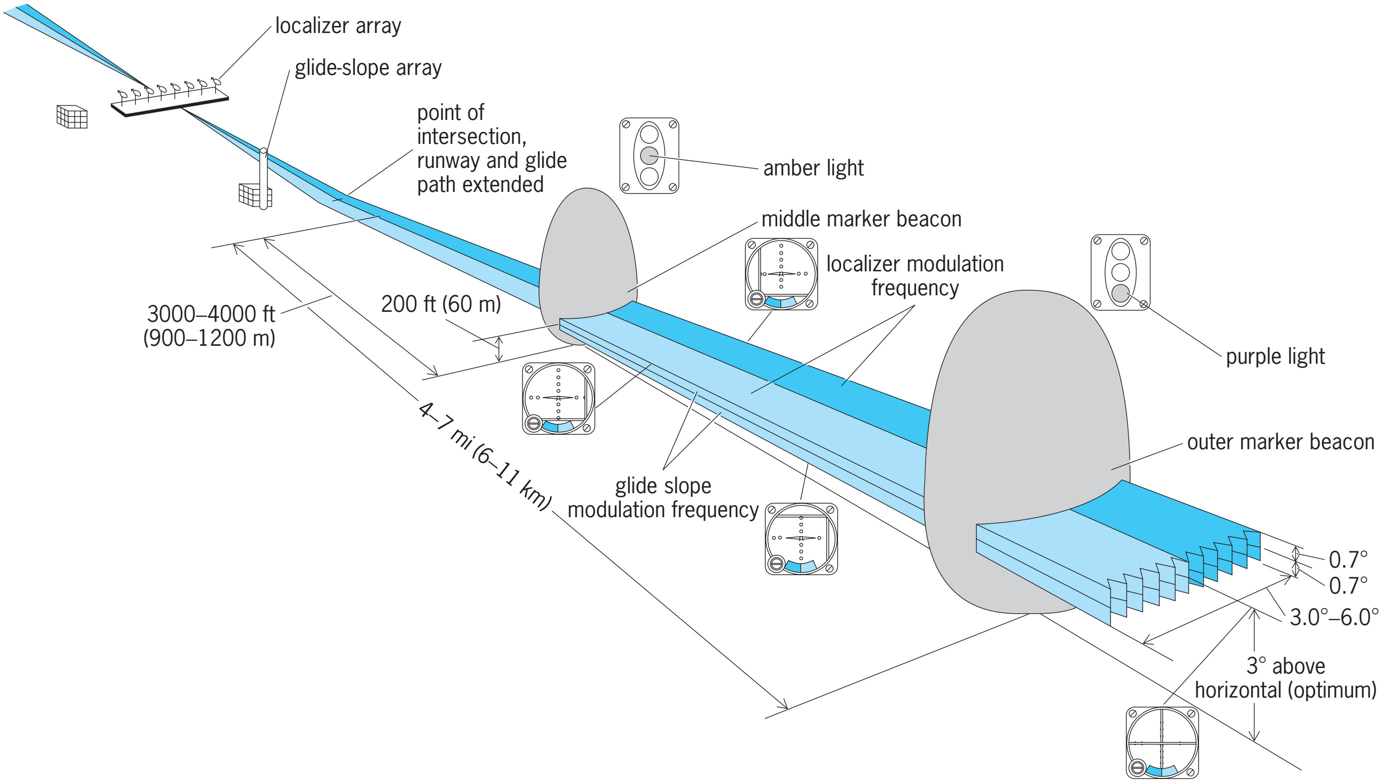


Fig. 1. Components of instrument landing system (ILS), and resulting signals used by the pilot. (*Federal Aviation Administration*)

special-purpose landing needs, but its widespread deployment is no longer planned.) *See* MICROWAVE LANDING SYSTEM (MLS); SATELLITE NAVIGATION SYSTEMS.

Standards. The International Civil Aviation Organization (ICAO) publishes the standards and recommended practices for the instrument landing system. The U.S. Federal Aviation Administration (FAA) publishes numerous manuals for the installation, maintenance, operation, monitoring, and calibration of the instrument landing system. A need for greater dependability of flight schedules has motivated the government to achieve even higher quality of ILS performance. Higher-quality ILS courses permit automatic flight controls to be used, especially when visibilities from the cockpit are measured to be only in hundreds of feet horizontally and tens of feet vertically. Another requirement for operation with lower visibilities is high system reliability. This reliability has been obtained largely through the use of solid-state transmitting equipment. All instrument landing systems are electronically monitored, and should out-of-tolerance conditions exist the transmitter is shut off automatically. *See* AUTOMATIC LANDING SYSTEM.

Spectrum availability. A limitation with ILS deployment is the availability of spectrum. Frequency assignments in some locations are difficult to make, and signals from the commercial frequency-modulation (FM) band, immediately below the ILS localizer band, have necessitated abandonment of some frequency channels. Relief has been obtained by making use of the directivity of the antenna arrays, and increasing the technical specifications for airborne receivers in order to make them more capable of rejecting undesirable signals. *See* RADIO SPECTRUM ALLOCATION.

Localizer. The localizer provides lateral guidance to the pilot by radiating its signal from an array of antennas usually placed about 600–1000 ft (183–305 m) beyond the stop end of the runway. There can be both a front and back course. The back course is an inexpensive means for providing guidance signals for approaches from the opposite direction when the wind direction dictates.

The localizer antenna array axis is perpendicular to the runway centerline. The horizontal extent and number of antennas in the array are determined by the need to confine the course guidance energy to narrow areas over the runway. Larger array dimensions are related inversely to the beam width of the signal in space. Narrower beams of energy prevent signal energy from being incident on buildings and then reflecting to cause course quality derogation. The dimensions of the array range from 40 to 130 ft (12 to 40 m) and the number of antennas from 6 to more than 20.

The three main types of localizer antenna are the V ring, a rather omnidirectional element; the traveling-wave types, which are unidirectional; and the log-periodic dipole (LPD) antenna, which is broad-band, usually unidirectional, and physically smaller (9 ft or 2.7 m long) and is located 6 ft (2.7 m) above ground. The LPD is the antenna in common use, typically in arrays of eight elements. *See* ANTENNA (ELECTROMAGNETISM).

Glide slope. The glide-slope array, located 300–400 ft (90–120 m) to the side of the desired touchdown point on the runway, in some respects can be considered a localizer on its side. There are five types of glide-slope arrays in use. Three are image systems; that is, they use the ground to form the path, and the space distribution patterns discussed below are the result of signals from the ground combining with direct signals from the antennas. The null-reference type, the simplest and most common of the image systems, uses two antennas 14 and 28 ft (4.3 and 8.6 m) above ground. The elements in the image arrays are three collinear dipoles backed by a corner reflector.

The performance of the image glide slope critically depends on the smoothness of the Earth's surface in front of the array out to 3 or 4 mi (4.8 or 6.4 km) into the approach region. A minimum of 1000 ft (305 m) of ground in front of the array with variations no greater than 12 in. (0.3 m) is desired. To overcome these limitations, special nonimage systems are used at approximately 20 sites. The two types of nonimage systems are the waveguide type and the endfire glide slope.

Theory. The localizer and glide-slope courses are formed by a technique called space modulation. A carrier frequency is generated in the transmitter, then amplitude-modulated with discrete 90-Hz and 150-Hz audio tones, thus giving sidebands 90 Hz and 150 Hz above and below the carrier. Some of the sideband energy is separated from the carrier by use of a passive, radio-frequency bridge. The resulting two types of energy, namely, carrier with sidebands (CSB) and sidebands only (SBO), are fed to the array such that the CSB main beam covers the general course area and SBO beams exist on either side of the centerline (**Fig. 2**). A sharp null in SBO energy is placed precisely on the centerline. The superposition of these signal patterns provides the space modulation that is interpreted by the airborne receiver as course guidance information.

In particular, the SBO lobes differ in phase by 180° on opposite sides of the null. In the case of the localizer (Fig. 2*a*), the 90-Hz sidebands of the CSB and SBO signals are in phase to the left of the centerline and therefore add, while the 150-Hz sidebands of these signals are 180° out of phase there and therefore subtract, thereby tending to cancel each other. To the right of the centerline the reverse is true: the 150-Hz sidebands are in phase and add, while the 90-Hz sidebands are 180° out of phase and subtract. Since only the CSB signal is received on the centerline, the 90-Hz and 150-Hz sidebands have equal amplitude there. The glide-slope theory is similar (Fig. 2*b*). The ILS receivers thus produce 90-Hz signals when the aircraft is above the glide-slope course and to the left of the localizer course, 150-Hz signals when the aircraft is below and to the right, and equal 90-Hz and 150-Hz signals when the aircraft is on course.

Receivers and display. ILS receivers are commonly double-conversion, superheterodyne types with amplitude-modulation (AM) detectors. The output of the detector is fed to 90-Hz and 150-Hz filters, after which the signal is rectified. In the most basic circuit this signal is used to drive zero-centered, dc microammeters. *See* AMPLITUDE-MODULATION DETECTOR; ELECTRIC FILTER; RADIO RECEIVER.

The receiver output is displayed in the cockpit on what is called a course deviation indicator (CDI). This instrument contains four basic elements: a vertical needle or pointer for displaying localizer course information, a horizontal needle for glide-slope information, and two flags which show when indications from the needles should be disregarded. (Marker beacon lights are usually on another display.) The needles deflect in a direction corresponding to the greater of the two detected modulations. The design and calibrations are for edge-to-edge deflection of the localizer needle to represent ±1.5 to 3.0° in azimuth depending on runway length, and for edge-to-edge deflection of the glide-slope needle to represent ±0.7° in elevation. For example, when the aircraft is left and below course, the pilot sees the needles deflected to the right and upward (Fig. 1) and should steer toward the needles to reach the on-courses. The on-courses turn out to be a vertical plane for the localizer and a cone open upward with a 184° vertex angle for the glide slope, making a 3° path angle for flight. In an aircraft flying on the localizer and glide-slope centerlines, the CDI shows centered needles, with the warning flags out of view.

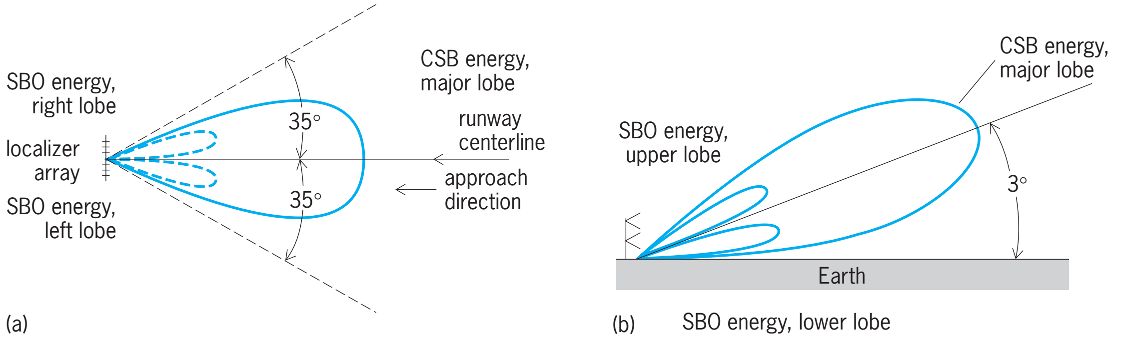


Fig. 2. Signal distribution space patterns for carrier-plus-sideband (CSB) energy and sideband-only (SBO) energies. (*a*) Localizer patterns in horizontal plane. (*b*) Glide-slope patterns in vertical plane. Angles are exaggerated.

Total localizer coverage is provided 35° on either side of the centerline. If narrower coverage is needed to protect against reflections, a two-frequency system is used and is based on the capture principle. The pilot need not be aware of any difference.

Marker beacons. The 75-MHz marker beacons are typically placed at two or three locations under the approach path on the extended runway centerline. They radiate a beam directly upward to indicate when the aircraft is at a precise location along the approach path. Because of the increase in the availability of other quality navigation information, the use of the markers is declining.

Sometimes associated with the instrument landing system is a low-frequency (200–400-kHz band), nondirectional radio beacon called a compass locator (COMLO), colocated with the outer marker. While very common and useful in the past, it too is finding less use, with other navigation signals supplanting it. *See* AIR NAVIGATION; ELECTRONIC NAVIGATION SYSTEMS. Richard H. McFarland

Bibliography. K. W. Bose, *Aviation Electronics*, 4th ed., 1990; Federal Aviation Administration, *Airman's Information Manual*, rev. ed. 1994; M. Kayton and W. Fried, *Avionics Navigation Systems*, 2d ed., 1997.

Instrument science

The systematically organized body of general concepts and principles underlying the design, analysis, and application of instruments and instrument systems. Instruments are devices for the acquisition and processing of information acquired from the physical world. They embrace, in particular, measuring and observation instruments.

Nature and scope. Instruments are very diverse in function and form. They differ according to measurands, range of magnitudes and the dynamic variations of the measurand, required accuracy, and nature and environment of application. Equally diverse is the range of technologies that can be used to realize a particular instrument function. To enable this information to be handled in such a way that it can be usefully applied to the design, analysis, or application of instruments, it must be organized on the basis of a systematic framework of general concepts and principles. This framework is termed instrument science. The term measurement science is also used for this field, when it is desired to emphasize measuring instruments and the measurement processes for which they are used.

The basis of instrument science is, first, the consideration of instruments as members of the general class of information machines; and second, an analysis and synthesis of the instruments as systems, applying the principles of system engineering.

Information machines. There exists a wide class of machines whose function is to acquire, process, and feed out information. This class includes measuring instruments, control apparatus, communication equipment, and computers. These machines function by the transformation of an input physical variable into an output variable in such a way that the output is functionally related to the input. Thus the output carries information about the input. This basic principle of functioning determines the general features of the analysis and synthesis of information machines. *See* COMPUTER; CONTROL SYSTEMS; ELECTRICAL COMMUNICATIONS.

Since information machines operate by maintaining a functional relation between a physical input and a physical output, they can be viewed as a material realization of a mathematical relation or model. Thus the analysis and synthesis of information machines relies on systematic methods of mathematical description of signals (information-carrying physical variables) and on mathematical modeling of signal transformation by machines. *See* INFORMATION THEORY.

More generally, the characteristics of the magnitude and time variations of a physical variable, which carry information about the objects and events of the real world, are termed symbols for those objects or events. The principles of symbolic representation form an important aspect of instrument science.

Instruments as systems. Instruments and indeed other forms of information machines are conveniently analyzed and synthesized as systems. A system is a set of interconnected components functioning as a unit. There is a set of general principles and techniques for treating complex entities as systems. It is known as the systems approach, and is based on the decomposition of the complex whole into individual components. Thus an instrument would be considered, for example, as an interconnection of a sensing element, a signal amplification element, a signal transmission element, a display, and so on.

The individual components are considered in the first instance in terms of function rather than form. Thus a particular instrument sensor is considered in the first instance as a displacement-to-voltage transducer rather than as a strain gage or a differential transformer.

Considering systems, such as instruments, as structures of functional building blocks makes clear that a very wide variety of instruments can be constructed from a much smaller variety of building blocks organized into a small variety of basic structures, such as a chain or loop connection. Further, it makes clear that apparently different systems—say, electrical and mechanical, or physical and biological—are essentially analogous.

An essential feature of the systems approach, following from its treatment of components of systems as functional blocks, is the use of mathematical models for systems analysis and design. Systematic methods of modeling are employed. The systems approach is further based on the application of systematic methods of design.

Finally, the systems approach to instruments and similar machines stresses the fact that instruments in their application are always embedded in a supersystem such as a control system or information system. It is the requirement of the supersystem which determines the design of the instrument. The instrument approach also stresses the significance of the environment in which the instrument is to operate, and the significance of all the phases of its life—from design and manufacture, through installation, commissioning, operation, and ultimate modification or disposal. *See* SYSTEMS ANALYSIS; SYSTEMS ENGINEERING.

Instrument system architecture. A typical measuring or observation instrument consists of a number of subsystems. First there is the system under measurement. This is connected to a sensor system and acts on the sensor by a flow of matter or energy. The sensor converts this flow into a signal, maintaining a functional relation between the input flow and the information-carrying characteristics of the signal. There is usually a signal-conditioning bloc, which converts this signal into a symbol. The symbol, in turn, may be conveniently handled by the following bloc, which performs any required functions of information transformation and communication. This system passes the information to the effector bloc, to further processing, or to the human operator. The measuring instrument system operates under the control of a control bloc. *See* TRANSDUCER.

An important part of the system is the human-machine interface. Through this interface, the operator effects supervisory control of the measurement process. The interface embodies also any displays. In the great majority of modern systems, and to a rapidly increasing extent, once information has been acquired by a sensor and conditioned, it is processed and effectuated by standard computing equipment. Thus, much of instrumentation is implemented by modern computer equipment. *See* DIGITAL COMPUTER; HUMAN-MACHINE SYSTEMS.

Models. An essential feature of instrument science is the development of systematic schemes for mathematical modeling or representation of the information-carrying physical variables or signals, and of the transformation performed by them on machines.

There are several levels of models of instruments employed: embodiment, power flow, signal flow, and information flow.

Embodiment models. These models relate the physical inputs and outputs of an instrument, or instrument component, in terms of features of its construction, that is, geometry and material properties.

Power-flow models. These models are based on considering power flow as a product of two conjugate variables, one an effort and one a flow. In mechanical systems they are force and velocity; in electrical systems, voltage and current; in fluid systems, pressure and flow; and in thermal systems, temperature and heat flow. All physical elements perform one of the following functions on power flow: storage, conversion, or transmission. The main power flow in elements may be modulated by controlling signals.

A modeling scheme based on these principles leads to a convenient classification of instrument elements and the recognition of analogies. Instrument elements such as sensors are basically one of three types: (1) power converters, such as a coil moving in a magnetic field, which transforms either a velocity into a voltage or a current into a force; (2) stores of two forms of energy, such as a moving-plate capacitor which stores mechanical and electrical energy and in which, for example, a mechanical velocity input produces an electrical voltage output; or, (3) a modulator, such as a resistance strain gage, in which the displacement straining the gage modulates the voltage-current relation of the gage.

Signal-flow models. Signal models are based on representing the signal as a time function (time domain) or as transforms such as Fourier and Laplace (frequency domain). The transformations performed by machines are modeled either in the time domain by convolution integrals or in the frequency domain by transfer functions. *See* FOURIER SERIES AND TRANSFORMS; LAPLACE TRANSFORM.

Information-flow models. The information level of modeling is concerned with information carrying characteristics of signals and their transformation. For example, in the case of digitally encoded information the signals are modeled in terms of digits, and their transformation in terms of truth tables and the like.

Signals. The principles underlying the analysis and design of signals, which carry information in information machines, and of signal transformation processes are a significant aspect of instrument science.

Intelligence. The processing of information in modern instruments is performed by subsystems of high information and knowledge processing power, which may be termed intelligent. The study of machine intelligence and the human cognitive processes which modern instruments emulate are a significant aspect of instrument science. *See* INTELLIGENT MACHINE.

Errors and uncertainty. The essential function of an instrument is the realization and maintenance of a prescribed functional transformation between input and output. In practice, the transformation cannot be realized or maintained perfectly. The result is error. Instrument science is concerned with providing a systematic framework for the description and analysis of errors.

This framework makes use of mathematical models of instruments. Errors are classified into two types, intrinsic errors and influence or disturbance errors. Intrinsic errors are the result of imperfections in the instrument input-output relations resulting, for example, from its imperfect determination or drift with time. Influence or disturbance errors are due to the effect of disturbing inputs acting on the instrument, such as changes of environmental temperature.

Errors may be determinate, that is, always the same for a particular value of input; or random, that is, different for different applications of a particular value of input. Random errors are modeled and analyzed statistically by assuming that they follow a gaussian or normal probability distribution. *See* DISTRIBUTION (PROBABILITY); PROBABILITY; STATISTICS.

Instrument design methodology. The great diversity of instrumentation means that instrument science cannot attempt to provide comprehensive catalogs of instrumentation. Rather it gives a systematic methodology of instrument design, based on the methodology of systems engineering, which enables appropriate instruments or instrument systems to be designed for a particular requirement.

The essential features of the methodology are primarily as pointed out above: the synthesis of instruments as systems of functional building blocks described by mathematical models. The other aspects are quantitative formulation of specifications in terms of objective criteria, the use of formal methods of optimization, and the application of formal techniques of decision in the consideration of alternative candidate designs. *See* DECISION THEORY; OPTIMIZATION.

Organization of technical knowledge. Knowledge about the working principles of instruments and instrument systems must be systematically classified and organized, so that it can be conveniently stored and retrieved. Three approaches may be adopted in classifying technical knowledge about instruments. The first is based on decomposing instruments into functional components and providing organized catalogs of such components and the principles of their combination. The second is organizing knowledge about measuring instrumentation on the basis of measurands. The third is organizing the knowledge about instrumentation on the basis of applications. L. Finkelstein

Bibliography. L. Finkelstein, Measurement and instrumentation science—an analytical review, *Measurement*, 14:3–14, 1994; L. Finkelstein and K. T. V. Grattan (eds.), *Concise Encyclopedia of Measurement and Instrumentation*, Pergamon Press, Oxford, 1994; P. H. Sydenham (ed.), *Handbook of Measurement Science*, 3 vols., 1982, 1983, 1992.

Instrument transformer

A device that serves as an input source of currents and voltages from an electric power system to instruments, relays, meters, and control devices. The basic design is that of a transformer with the primary winding connected to the power system, and the secondary winding to the sensing and measuring equipment. Data from these devices are necessary for the operation, control, and protection of the power system. The primary reason for setting up the instrument-transformer interface is to provide analog information at low voltage levels, insulated from the higher system voltages. The range of use is from 480 V through the maxima of the 765–1000-kV power systems. *See* ELECTRIC POWER SYSTEMS; TRANSFORMER.

Current transformers are connected in series with the power conductor. In many cases this conductor serves as the one-turn primary. The principal types are the window type, where the power conductor is passed through a hole in the center of the current transformer; the bar type, where the power conductor is fastened to a one-turn bar which is part of the current transformer; and the bushing type, a toroidal core and winding that is slipped over the insulating bushings of circuit breakers, transformers, and so forth. At voltages of 345 kV and above, the current transformer may be mounted on top of a tall porcelain insulator. The standard secondary rating is 5 A, except for a few 1-A current transformers used for relaying, mostly outside the United States.

Voltage transformers and coupling capacitor voltage transformers are connected in parallel from one conductor to another or to ground. The coupling capacitor voltage transformer is widely used at the higher system voltages of 115 kV and above. It is a voltage transformer tapped across part of a capacitor unit connected from the conductor to ground. The standard secondary ratings are 120 (115) V line to line, or 69.3 (66.5) V line to neutral.

There are other means of providing primary system information to the secondary equipment, involving electronics, light pipes, optical fibers, and so forth. Some of these are in service on power systems. They are used primarily for instrumentation and data acquisition.

A typical equivalent circuit diagram for these transformers is shown in **Fig. 1**. The objective of the design of instrument transformers is to have the primary quantities precisely reproduced in the secondary, in both magnitude and phase angle, and under all steady-state and transient conditions. In practice, this goal can only be approached since the iron in the transformer core must be magnetized (by the exciting impedance and current, Z_e and I_e, in Fig. 1) resulting in losses through the device. The actual secondary current available I_s is equal to the secondary value I_r of the primary current I_p minus the exciting current I_e. Air-core (linear coupler) transformers, which do not suffer from these losses, are employed in relaying, but their use is quite limited.

Accuracy requirements. For instrumentation and especially for metering used for billing and revenue, very high accuracy is important for load current (which ranges from 0 to 5 A in the secondary) and load voltages (which range from 90 to 110% of the rated value). On the other hand, for protective relaying, reasonable ratio accuracy is important over a

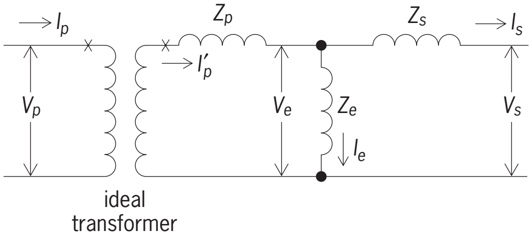


Fig. 1. Equivalent diagram for instrument transformers. The ideal transformer is required to change volts and amperes from primary to secondary. With per-unit values, the ideal transformer ratio is 1:1, and so for simplification is normally omitted. The ideal transformer has the ratio 1:*n* for current transformers and *n*:1 for voltage transformers, where *n* is the number of turns.

very wide range from just above zero to a maximum of 100 to 300 A in the secondary. Phase-angle error as such is not directly a problem in relaying. Thus, particularly for current transformers, metering and relaying designs are quite different and are not interchangeable. For voltage transformers, the values of impedance in the primary and secondary, Z_p and Z_s, are kept low to minimize voltage loss and phase shift from the primary to the secondary. For current transformers, Z_e is kept high to minimize current loss from primary to secondary. Z_p is negligible when the primary has one or a small number of turns.

Load. The equipment (load) connected to the instrument transformer secondary is commonly called the burden Z_B and is often expressed in volt-amperes (VA_B) for each device. Thus Eq. (1) is valid, where

$$Z_B = \frac{\mathrm{VA}_B}{I^2} \quad \text{ohms} \qquad (1)$$

I is the current in amperes at which the burden is measured. The total load Z_L is the sum of all the connected equipment, including the connecting leads. The phase angle is important, especially for metering.

Performance standards. The performance of instrument transformers is standardized in the United States by American National Standards Institute Standard C57.13 for both metering and relaying service. Accuracy classes are specified. For metering, transformer correction factors for standard burdens provide the means to determine performance for both current transformers and voltage transformers. For relaying, current transformer performance can be determined by several methods.

Current transformer. For current transformers connected in the three phases of a three-phase power system, the current transformer ratio is selected by determining the maximum current that can flow for any normal or emergency operating condition. A current transformer ratio R_c is selected so that this current, divided by the ratio, is around 4 but less than 5 A. Most indicating instruments used have 5-A movements regardless of the scale, so this avoids pegging the indicator. Also, much of the equipment to be connected has a 5-A rating.

For relaying, the next step is to determine the maximum fault current I_F at which the relays must operate. Assume that the current transformer will pass this current through the total burden, including the current transformer secondary impedance Z_s with negligible exciting current I_e. To accomplish this, the current transformer must develop an exciting voltage V_e given by Eq. (2), where Z_s has a value of

$$V_e = \frac{I_F}{R_c}\,(Z_s + Z_{B,\text{total}}) \qquad (2)$$

0.0025 ohm per turn.

There are two types of current transformers used in relaying: ANSI class C, where the leakage flux in the core does not significantly affect the ratio (Z_s is small); and ANSI class T, where the leakage flux has appreciable effect on the ratio (Z_s is significant). Many relaying current transformers are class C, and their performance can be calculated and evaluated from excitation curves supplied by the manufacturer or determined by test. Typical curves for a multiratio-bushing-type current transformer are shown in **Fig. 2**. The value of V_e, determined from

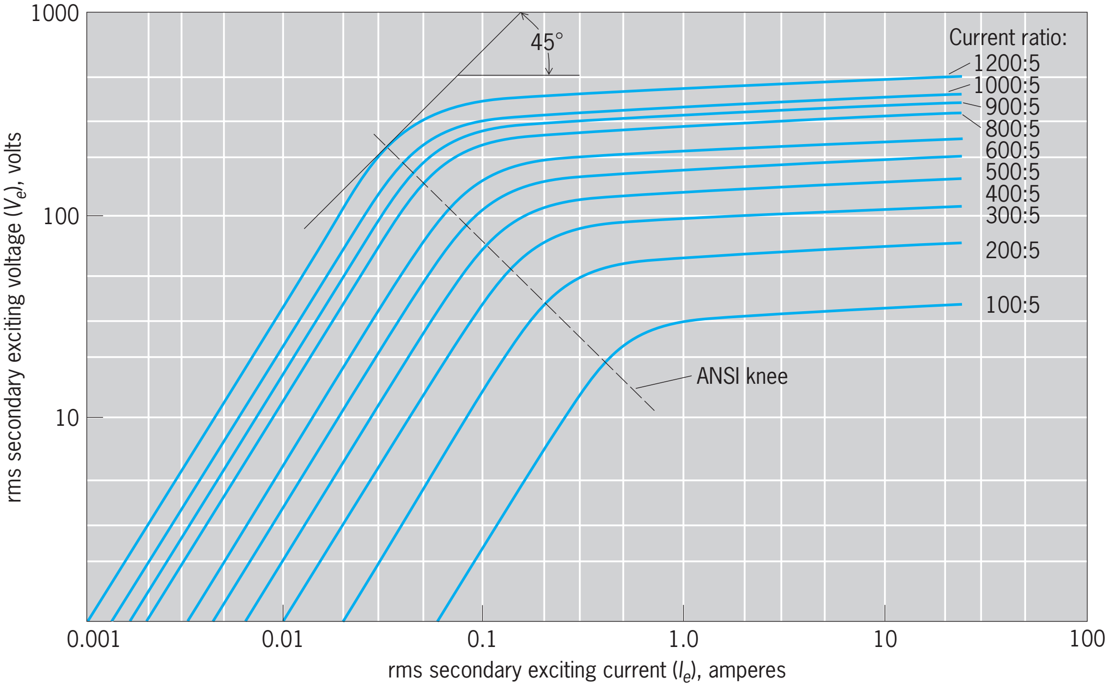


Fig. 2. Typical excitation curves for a multiratio class C current transformer. Below the line labeled ANSI knee, the exciting current for a given voltage for any unit will not exceed the curve value by more than 25%. Above this line, the voltage for a given exciting current for any unit will not be less than 95% of the curve value. (*After American National Standards Institute, Requirements for Instrument Transformers, ANSI/IEEE Stand. C57.13, 1978*)

Standard burdens for current transformers with 5-A secondaries*

ANSI burden designation	Impedance, ohm	Volt-amperes at 5 A	Power factor	Service
B-0.1	0.1	2.5	0.9	Metering
B-0.2	0.2	5.0	0.9	Metering
B-0.5	0.5	12.5	0.9	Metering
B-0.9	0.9	22.5	0.9	Metering
B-1.8	1.8	45.0	0.9	Metering
B-1	1.0	25.0	0.5	Relaying
B-2	2.0	50.0	0.5	Relaying
B-4	4.0	100.0	0.5	Relaying
B-8	8.0	200.0	0.5	Relaying

*After American National Standards Institute, *Requirements for Instrument Transformers,* ANSI/IEEE Standard C57.13, 1978.

Eq. (2), provides the value of I_e required. If, for the maximum fault, V_e for the connected burden is below the knee of the curves, I_e will be small and I_s essentially equal to the primary current divided by R_c.

For relaying class T current transformers, overcurrent ratio curves are supplied (**Fig. 3**) in terms of standard burdens connected to the terminals of the current transformer. The **table** shows these burdens. From the total burden connected, the current transformer performance can be evaluated from these curves.

As an example, consider a current transformer with a ratio of 800:5 and an external burden of 1.6 ohms. The maximum fault is 10,000 A primary. If the transformer is class C, V_e is given by Eq. (3),

$$V_e = \frac{10,000}{160}(0.4 + 1.6) = 125 \text{ V} \qquad (3)$$

where, from Fig. 2, I_e is approximately 0.045 A, quite small for the 62.5 A secondary for the fault. If the current transformer is type T, the burden for Fig. 3 is 1.6 ohms, and the ratio is essentially unity for the fault current of 62.5/5 in the secondary or 10,000/800 in the primary, both equal to 12.5 on the curve.

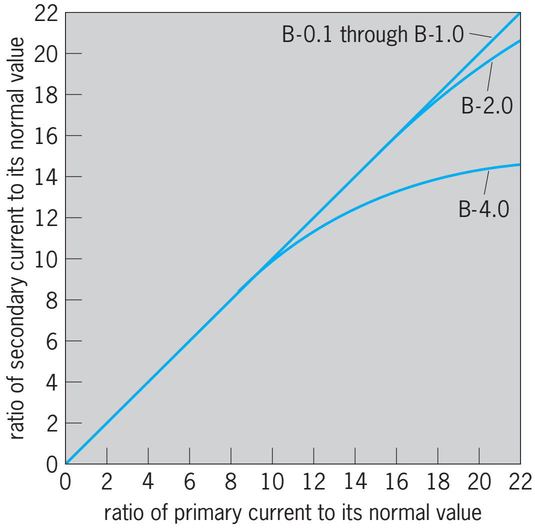


Fig. 3. **Typical overcurrent ratio curves for class T current transformers. (*After American National Standards Institute, Requirements for Instrument Transformers, ANSI/IEEE Stand. C57.13, 1978*)**

In many cases the ANSI accuracy class designation can be used without detailed information on the current transformer. In relaying, classes C and T are followed by a number indicating the voltage that the transformer will deliver to a standard burden at 20 times rated secondary current without exceeding 10% ratio error. This voltage is the terminal voltage V_s, and for multiratio current transformers applies at the full winding tap. Proportionately lower voltages are developed at the other taps. In the example above where $V_s = 62.5 \times 1.6 = 100$ V, any current transformer rated C100 or T100 or higher could be used.

Transient performance. Transient performance of instrument transformers is a complex subject, not covered in the present standards. In general, modern transformers transform the various harmonics and wave shapes very adequately except for the direct-current decrement. Fault currents usually are offset by various amounts. This occurs because at the instant of the fault the flux relations in the circuit are not correct for the sudden change in current. As a result, a dc component is produced which then decays according to the inductance/resistance (L/R) ratio of the power system. This is a random effect so that the dc peak is variable and may or may not reach its maximum possible value. This direct current inhibits current transformer performance by saturating the core so that I_e becomes very large, and I_s is reduced. It is not practicable to build a current transformer to minimize this effect, so it must be managed by relay design and application to avoid incorrect operations. As a generality, with the maximum dc offset, a current transformer will reproduce about a half-cycle or so of the wave before saturation reduces the secondary current to very low values. As the direct current decays, the current transformer recovers to eventually accurately reproduce the primary current. Thus relays operating in the order of the half-cycle may operate before saturation, while slower relays will be delayed. In differential-type protection where several current transformers are interconnected and must operate as a team, application rules are established to prevent misoperations.

Voltage transformers. Voltage transformers are seldom a problem in performance and in capacity as long as the connected burden is within the rating of the units. In three-phase power systems, the common connections are three units connected wye-grounded-wye-grounded or two units in open delta across the three phases. The connected load is generally line to line. Special applications may require only one voltage transformer connected usually line to ground. Coupling capacitor voltage transformers and some voltage transformers have two secondaries. In such cases the primaries are connected wye-grounded. One secondary is connected wye-grounded to provide three-phase voltages for relays, instruments, and so forth. The other windings are connected in delta with one corner open (broken delta) and connected to ground relays.

Faults reduce the voltage, and power system overvoltage is carefully controlled so that coupling capacitor voltage transformers and voltage transformers operate only from very low voltages to around 100% of their ratings, so that saturation is not a problem. *See* ELECTRICAL MEASUREMENTS. J. Lewis Blackburn

Bibliography. American National Standards Institute, *Requirements for Instrument Transformers*, ANSI/IEEE Standard C57.13, 1978; J. S. Blackburn, *Protective Relaying*, *Principles and Applications*, 2d ed., 1997; W. A. Elmore (ed.), *Protective Relaying Theory and Practice*, 1994; Westinghouse Electric Corp., *Applied Protective Relaying*, 1982.

Instrumental conditioning

Learning based upon the consequences of behavior. For example, a rat may learn to press a lever when this action produces food. Instrumental or operant behavior is the behavior by which an organism changes its environment. The particular instances of behavior that produce consequences are called responses. Classes of responses having characteristic consequences are called operant classes; responses in an operant class operate on (act upon) the environment. For example, a rat's lever-press responses may include pressing with the left or with the right paw, sitting on the lever, and so on, but all these taken together constitute the operant class called lever pressing.

History. Instrumental conditioning was first studied systematically by Edward L. Thorndike in the late nineteenth century. Thorndike constructed puzzle boxes in which he placed cats and other animals. The animal could get out of the box (and, usually, get fed) by operating a latch or some other device that released the door. Thorndike found that the time it took the animal to get out of a particular box typically decreased gradually over many trials. He spoke of this as the law of effect; behavior was "stamped in" by the effects it had or, in other words, by its consequences. Instrumental learning in these types of situations was sometimes called trial-and-error learning.

Subsequently, instrumental learning was studied with many different species in many devices, such as mazes and runways. Experimenters asked questions such as whether this learning took place in discrete steps or continuously, and whether organisms learned movements (response learning) or properties of the environment (stimulus learning). The experiments had the common result that animal response became more likely when it had certain consequences. Graphs showing how behavior changed during an experiment were called learning curves: these plot time to respond as a function of number of trials, percentage of correct responses, or proportion of animals performing successfully. But the shapes of the curves depended so much on differences in experimental detail that the progress of learning could not be defined in any consistent way.

The performances of the animals were too complicated to be clearly interpreted. For example, measuring the time it took a rat to get through a maze showed little of how learning proceeded at a single choice point. An important innovation was B. F. Skinner's design of a chamber in which an animal could repeatedly emit some easily specified response without the experimenter's intervention (the response was said to be emitted because it was not produced by any particular stimulus). A chamber typically included lamps and speakers for presenting stimuli, switches attached to response devices such as levers or disks on the chamber wall, and dispensers that presented food or water as consequences for responses. These chambers were used to demonstrate reinforcement (an increase in behavior that occurs because of the consequences of responses), extinction (the return of behavior to its previous lower levels when these consequences are discontinued), and the effects of schedules of reinforcement (arrangements of intermittent consequences for responses, as, for example, after every tenth response or on the average once every 2 min). The relations between responses and their consequences are called contingencies.

Reinforcers and punishers. Consequences of responding that produce increases in behavior are called reinforcers. For example, if an animal has not eaten recently, food is likely to reinforce many of its responses. The effectiveness of a reinforcer depends on the response that is reinforced. Suppose that a rat is deprived of water but has had unlimited access to a running wheel. If access to water is made a consequence of running in the wheel, the opportunity to drink will reinforce wheel running. But suppose next that the rat is given unlimited access to water but is prevented from running in the wheel. If access to the wheel is made a consequence of drinking, the opportunity to run will now reinforce drinking. In other words, the reinforcement relation is reversible. Whether a particular event will reinforce a response or not depends on the relation between the response and the behavior for which its consequences provide an opportunity. Consequences of responding that are effective as reinforcers may include many kinds of events: simple sensory events like making contact with something as a result of reaching for it or seeing something that one has turned to look at; biologically significant events such as sexual contact or the opportunity to eat or drink; and the removal or prevention of something, as in escape from or avoidance of painful stimuli.

Some consequences of responding called punishers produce decreases in behavior. The properties of punishment are similar to those of reinforcement, except for the difference in direction. The effects of reinforcement are temporary in that responding decreases to earlier levels when reinforcement is discontinued (extinction); the effects of punishment are also temporary in that punished responding returns to its earlier higher levels when punishment is discontinued. As with reinforcers, many types of events can serve as punishers, and their effects depend on their relation to the response that is punished. For example, if a painful stimulus makes a rat jump, the stimulus will probably not be an effective punisher of jumping, though it might work for other responses such as grooming.

Discriminative or occasion-setting stimuli. The consequences of its behavior are perhaps the most important properties of the world about which an organism can learn, but few consequences are independent of other circumstances. Organisms learn that their responses have one consequence in one setting and different consequences in another (stepping on the gas, for example, has some consequences when the traffic light is green that differ from those when the traffic light is red). Stimuli that set the occasion on which responses have different consequences are called discriminative stimuli (these stimuli do not elicit responses; their functions are different from those that are simply followed by other stimuli). When organisms learn that responses have consequences in the presence of one but not another stimulus, their responses are said to be under stimulus control. For example, if a rat's lever presses produce food when a light is on but not when it is off and the rat comes to press only when the light is on, the rat's presses are said to be under the stimulus control of the light. The relations among a discriminative stimulus, a response, and a consequence have been described by Skinner as a three-term contingency, in that all three terms must enter into what the organism learns: in a particular "environmental setting," its "behavior" produces certain "outcomes." These contingencies do not explain instrumental learning; they are what the organism learns. *See* CONDITIONED REFLEX; LEARNING MECHANISMS. A. Charles Catania

Bibliography. A. C. Catania, *Learning*, 4th ed., 1997; M. Sidman, *Tactics of Scientific Research: Evaluating Experimental Data in Psychology*, 1988; B. F. Skinner, *The Behavior of Organisms*, 1938; B. F. Skinner, *Contingencies of Reinforcement*, 1969; E. L. Thorndike, *Animal Intelligence*, 1911.

Instrumentation amplifier

A special-purpose linear amplifier, used for the accurate amplification of the difference between two (often small) voltages, often in the presence of much larger common-mode voltages. An instrumentation amplifier has a pair of differential (usually high-impedance) input terminals, connected to sources V_{in1} and V_{in2}; a well-defined differential-mode gain A_{DM}; and a voltage output V_{out}, satisfying the relationship given in Eq. (1). It differs from an operational

$$V_{\text{out}} = A_{\text{DM}}(V_{\text{in1}} - V_{\text{in2}}) \qquad (1)$$

amplifier (op-amp), which ideally has infinite open-loop gain and must be used in conjunction with external elements to define the closed-loop transfer function. At one time instrumentation amplifiers were built in discrete or hybrid form using operational amplifier and resistor networks, but they are now readily available as inexpensive monolithic integrated circuits. Typical commercial amplifiers provide preset gains of 1, 10, 100, and 1000. In some cases, the gain may be set to a special value by one or more external resistors. The frequency response invariably is flat, extending from 0 (dc) to an upper frequency of about 1 kHz to 1 MHz. *See* INTEGRATED CIRCUITS; OPERATIONAL AMPLIFIER; RESISTOR.

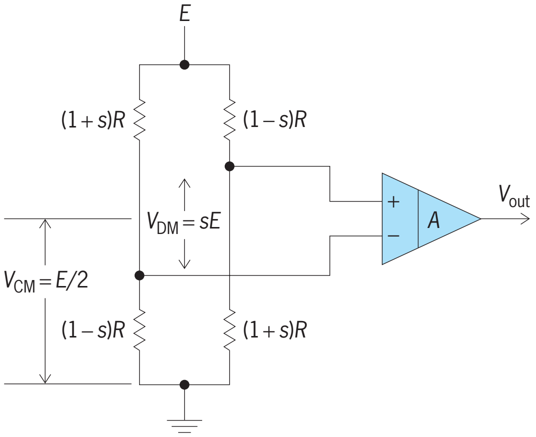


Fig. 1. Strain-gage bridge which produces a small wanted signal on a large bias, connected to an instrumentation amplifier that responds only to the wanted signal. The amplifier symbol indicates that the differential gain is controlled, unlike that of an operational amplifier.

Instrumentation amplifiers are used to interface low-level devices, such as strain gages, pressure transducers, and Hall-effect magnetic sensors, into a subsequent high-level process, such as analog-to-digital conversion. **Figure 1** shows a typical strain-gage bridge, which generates a small differential-mode output V_{DM}, proportional to the strain factor *s*, superimposed on a large common-mode voltage V_{CM}. The desired output level might be several volts, and is typically ±10 V full-scale. *See* HALL EFFECT; PRESSURE TRANSDUCER; STRAIN GAGE.

An important requirement of an instrumentation amplifier is that it responds only to the difference signal and rejects other signal components. A real amplifier has some finite response to its common-mode input, as defined in Eq. (2). The common-mode

$$V_{\text{CM}} = \frac{V_{\text{in1}} + V_{\text{in2}}}{2} \qquad (2)$$

rejection ratio (CMRR) is a measure of how well the amplifier performs this rejection. If A_{CM} is the common-mode gain, Eq. (3) gives the common-mode

$$\text{CMRR (dB)} = 20 \log_{10} \frac{A_{\text{DM}}}{A_{\text{CM}}} \tag{3}$$

rejection ratio, conventionally expressed in decibels.

At low frequencies (0 to a few kilohertz), common-mode rejection ratios between 90 and 120 dB are possible. At higher frequencies, the common-mode rejection ratio is eroded by small imbalances in the differential signal path. For example, differences in the source impedances may result in the conversion of a common-mode signal to a differential-mode one. Many effects in the amplifier cause degradation of the common-mode rejection ratio at high frequencies.

A classical topology uses two operational amplifiers as voltage followers (to provide the required high-input impedance) and a third operational amplifier in a differencing configuration (**Fig. 2***a*). The output voltage is given by Eq. (4), so that the differential-

$$V_{\text{out}} = \frac{R_2}{R_1}(V_{\text{in1}} - V_{\text{in2}}) \tag{4}$$

mode gain is simply the resistor ratio R_2/R_1. If like resistors match perfectly, the common-mode rejection ratio is limited only by the active elements. However, a very high degree of resistor matching is required in this topology to achieve high rejection of the common-mode signal; for example, a 0.1% mismatch would limit the common-mode rejection ratio to 60 dB.

A more accurate topology is used in many high-performance instrumentation amplifiers (Fig. 2*b*), with output voltage given by Eq. (5). By raising the

$$V_{\text{out}} = \left(\frac{2R_2}{R_1} + 1\right)(V_{\text{in1}} - V_{\text{in2}}) \tag{5}$$

gain ahead of the differencing amplifier (now a unity-gain structure), the effect of all resistor mismatches is minimized, and the amplifier has a high common-mode rejection ratio even when resistor matching is moderately poor, particularly at high gains. Many other configurations have been used with varying success. As with conventional operational amplifiers, modern monolithic instrumentation amplifiers have low input offset voltage (as low as 50 microvolts), low input-bias current (0.1–50 nanoamperes), and low voltage and current noise. Laser trimming of the integral resistors is often employed to achieve very high-gain accuracy. *See* AMPLIFIER; DIFFERENTIAL AMPLIFIER.

Barrie Gilbert

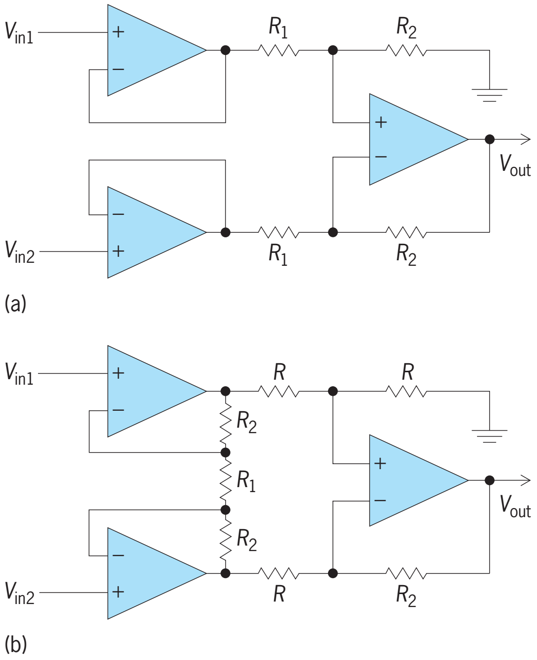


Fig. 2. Instrumentation amplifier topologies. (*a*) Simple topology based on three operational amplifiers. (*b*) Improved topology used in a modern integrated-circuit instrumentation amplifier.

Bibliography. E. W. Greeneich, *Analog Integrated Circuits*, 1996; J. C. C. Nelson, *Operational Amplifier Circuits: Analysis and Design*, 1995; D. H. Sheingold (ed.), *Transducer Interfacing Handbook*, 1980.

Instrumented buoys

Unattended, floating structures equipped with systems for collecting, processing, and transmitting meteorological and oceanographic data. Such information is useful for many purposes, including storm warnings and forecasts, coastal engineering, climatology, and oceanographic and atmospheric research.

In earlier times, merchant vessels supplied almost all our knowledge of ocean weather and climate. Ships continue to play a major role in reporting weather and surface conditions (waves, swell, and temperature), but the irregular and sporadic nature of these observations, especially in the sparsely traveled Southern Hemisphere, makes it difficult to obtain a systematic picture of the state of the atmosphere and how it changes with time. Apart from crude estimates of surface currents through dead reckoning, ships in transit tell us nothing about the ocean. So, almost all our knowledge of ocean currents and their variability today comes from instrumented buoys, which measure ocean currents and related properties, particularly temperature. Moored systems, anchored to the ocean bottom, record currents and water properties as the flow passes the buoy (the eulerian framework). Drifting buoys move with the waters, indicating where they go (the lagrangian framework). In both cases, additional sensors may record a variety of parameters, such as temperature, pressure, acceleration, and water properties.

Moored systems. Surface moorings come in many shapes and sizes depending upon a number of factors, such as research versus operational requirements, duration of deployment, surface versus subsurface measurements, strength of currents, and weather considerations. To operate for a long time,

Fig. 1. Atlas toroid buoy with its meteorological instrumentation. (*Pacific Marine and Environmental Laboratory, NOAA*)

moored buoys require highly reliable components, including a strong mooring line to prevent failure due to wear and tear in heavy seas. This in turn requires large buoys with substantial surface flotation to support the weight of the line. These buoys support a suite of atmospheric sensors and power systems for their operation and data telemetry.

Surface buoys have a long history, particularly along the United States Atlantic, Gulf, and Pacific coasts in reporting offshore weather and sea surface conditions for shipping, fishing, and recreational use. Their size and shape vary enormously, depending upon their application. Doughnut-shaped toroid, discus, and boat-shaped NOMAD (Navy Oceanographic and Meteorological Automatic Device) buoys can survive very harsh weather and current conditions (**Fig. 1**). In response to the impact of the equatorial Pacific on year-to-year variations in the Earth's weather (El Niño-La Niña), an extensive array of moored toroids have been deployed across the equatorial Pacific (about 70) and Atlantic (about 10) oceans. They transmit weather and water temperature from the surface down through the shallow thermocline (the boundary between the warm upper and cold deeper waters) to 500 m. The weather service uses the real-time data sent from these buoys for their forecasting activities. The public can also access the data via Internet websites. To ensure reliable performance, the buoys get serviced or replaced every year. *See* EL NIÑO; METEOROLOGICAL INSTRUMENTATION; TELEMETERING; WEATHER FORECASTING AND PREDICTION.

Studies of the subsurface ocean typically employ moorings with flotation below the surface. Subsurface moorings experience much less fatigue of the hardware and mooring line due to the absence of wave motion, permitting the use of lighter-weight, less expensive hardware. More importantly, the absence of surface wave motion results in a much calmer mooring line, greatly facilitating accurate current measurements. Subsurface moorings also eliminate the risk of piracy and entanglement with fishing nets, a major source of equipment and data loss for surface moorings. However, these moorings require costly subsurface flotation devices because they must have the strength to withstand the higher hydrostatic pressure. Large steel spheres are used for near-surface flotation applications, and glass balls are used at greater pressures. Syntactic foam, consisting of a matrix of glass microspheres embedded in epoxy, has proven very effective in applications where the flotation must have a special shape to support instruments, yet remain streamlined to minimize drag forces in strong currents.

A tall mooring reaching close to the surface in deep water, as in the Gulf Stream, experiences very strong drag forces due to current flow past the wire, the weight of the instrumentation, and the weight of the flotation device at the top of the mooring. In the absence of sufficient flotation, these drag forces can pull the mooring over many hundreds of meters and can exert a very large drag force on the anchor, which therefore must be very heavy (typically an assembly of discarded railroad wheels weighing 1–3 tons). Subsurface moorings require considerable design attention because the expected drag forces determine the required flotation, the minimum necessary strength for the mooring wire, and the diameter of the wire. The mooring line typically consists of a jacketed steel wire for strength and for protection against fish bite in the upper ocean, and lighter-weight synthetic fiber-rope at greater depths.

For many years the use of subsurface flotation precluded the possibility of real-time telemetry since radio waves cannot pass through the electrically conducting seawater. New technologies avoid this limitation if the top of the mooring reaches close to the surface. In one design, a small winch at the top of the mooring sends a small transmitter module to the surface that transmits its data to satellites, and then pulls itself back below the surface. In another design, a small float rides the surface but remains attached to the top of the mooring and its flotation device by means of a stiff yet flexible tether known as an S-tether. The stiffness prevents kinking, while the flexibility lets the small surface buoy with its embedded radio transmitter follow the surface motion without affecting the rest of the mooring.

A novel mooring design called the moored profiler consists of a taut subsurface mooring along which a module drives itself up and down, recording water properties and currents between the bottom and near the surface. The profiler can also send the data to the top of the mooring inductively via a magnetic coupling between the profiler and the wire. Using an S-tether or other surface module, the system can telemeter the data with little delay for analysis. Real-time telemetry also means data are available up to the point of failure.

Drifting buoys. These buoys drift with the waters, either as surface drifters or as subsurface floats. In order to follow the waters while minimizing the effect of winds, surface drifters have very little exposure at the surface but have large drag elements hanging beneath. Today, the nearly universal drifter design includes a spherical flotation element and a long, large-diameter canvas tube with large holes, known as a holey sock. Its large size guarantees that the drifter moves with the waters around the sock, readily pulling the small surface sphere with it. Depending upon the application, the sock might hang just to 10 m or as deep as 100 m. The movement of the drifters are tracked by satellites. These drifters have made major contributions to our knowledge of the surface circulation of all the oceans. Some new designs come equipped with temperature sensors, and increasingly, with barometric pressure gauges. They also telemeter valuable sea surface weather information, including acoustic measurements of winds. These buoys cost-effectively complement the functions of the surface buoys.

Subsurface drifters also have a long history, beginning with the development of the neutrally buoyant float equipped with a small pinger (a sound source) so that a ship can follow it as it drifts. The floats achieve their neutral buoyancy at depth by adjusting their weight very carefully to match that of the displaced waters (Archimedes' principle). A float must also have less compressibility than water for it to remain at depth. Suppose, for example, a disturbance such as an internal wave deflects the float upward. The float will expand less than the surrounding waters due to its compressive stiffness, making it slightly heavier than the surrounding waters, causing it to sink back toward its equilibrium depth. These early float-based studies revealed unexpectedly energetic subsurface currents. *See* ARCHIMEDES' PRINCIPLE.

The acoustic transparency of the ocean, especially at low frequencies (a few hundred hertz), led to the development of low-frequency sound sources, enabling acoustic tracking over more than 2000 km though the deep sound or SOFAR (sound fixing and ranging) channel. Earlier float designs had the floats transmitting signals on a precisely timed basis to two or more underwater listening stations or hydrophones. Given the speed of sound in water, one can estimate the position of the float to a few kilometers' accuracy, and hence its drift as a function of time. Today, acoustic tracking works in the opposite direction: moored sound sources function as an underwater navigation system for drifting floats, which listen to and record the arrival times of the signals. At the end of a float's underwater mission, it surfaces and telemeters the stored data via satellite to the users, who can then reconstruct the trajectory of the float.

In addition to acoustic tracking, floats always measure temperature and pressure. Recent float designs include oxygen sensors. These RAFOS (SOFAR spelled backward) floats can either float at constant depth or follow density surfaces as they change depth. In the latter case, the floats incorporate a compressible element to give them the same compressibility as that of seawater, freeing them to follow the waters in the vertical. To ensure that they stay with the same waters, the floats have virtually no coefficient of thermal expansion (they are made of borosilicate glass). Using the same argument as above with respect to pressure, imagine a float temporarily displaced upward into warmer waters. The lack of thermal expansion compared to that of seawater makes it slightly heavier than the surrounding waters, forcing it back to the water temperature it came from. More generally, it follows that once the float has settled on a given density surface, it will stay

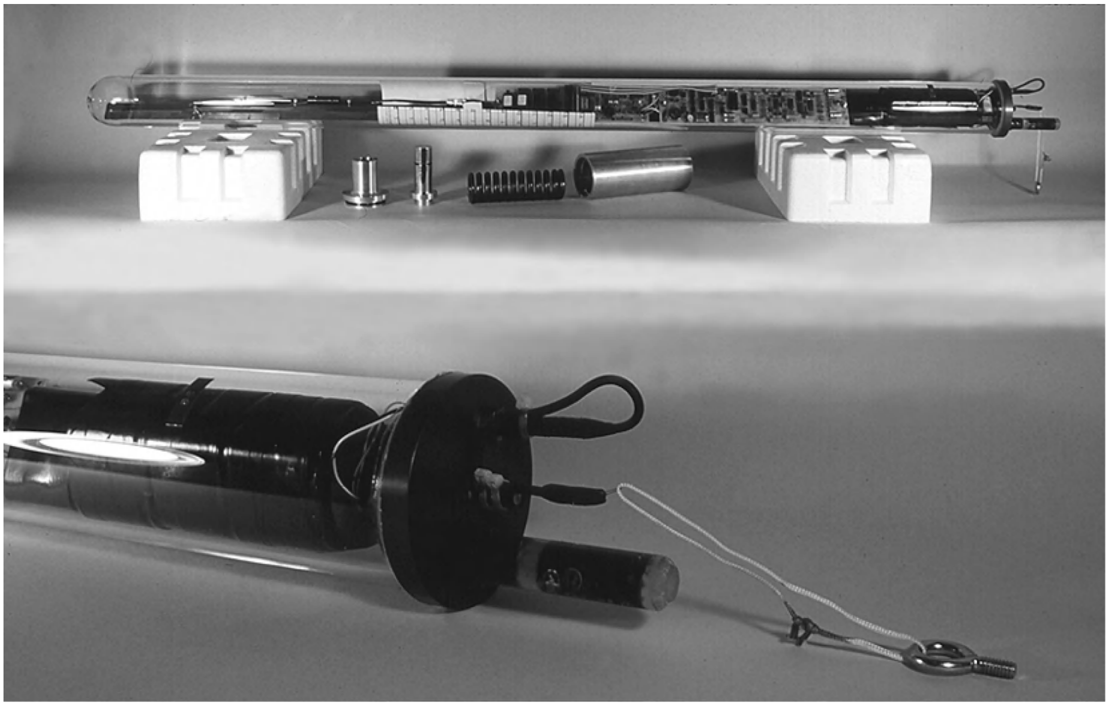


Fig. 2. RAFOS float showing the whole instrument, with a detail of the right end in the foreground. (*Graduate School of Oceanography, University of Rhode Island*)

with that surface regardless of temperature and salinity, since neither can affect the density of the float. This feature confers major advantages for the study of dispersion and mixing processes in the ocean. These acoustically tracked floats have yielded much insight into the horizontal movements of the ocean over a wide range of space and time scales and at all depths (**Fig. 2**).

Another class of floats focuses on the vertical structure of the ocean. Drifting at some preset depth (typically 1000–1500 m), they ascend to the surface while sampling pressure, temperature, and salinity. While at the surface, the floats use the Global Positioning System (GPS) to determine their position, which they broadcast, along with the profile data, to a satellite. Alternatively, some satellites can determine their positions. The floats then return to depth and continue their drift for a predetermined period (typically a couple of weeks) before repeating the process. Their change in position between surface reports gives valuable information on the subsurface circulation. Satellite navigation allows their deployment anywhere in the global ocean. The combination of wide-area deployment of these profiling floats and telemetry provides a very effective means for obtaining wide-area information on the interior state of the ocean in real time (**Fig. 3**). Advanced planning calls for their deployment around the world in large numbers as part of the Global Ocean Observing System. *See* NAVIGATION; SATELLITE NAVIGATION SYSTEMS; SOFAR; SONAR; UNDERWATER SOUND.

Sensors. Buoy sensors can be categorized in many ways. For example, they can be grouped according to whether they measure scalar or vector properties. Scalar sensors measure the state of a fluid, such as temperature, pressure, humidity, salinity, and light intensity. Vector sensors measure speed and direction of winds and currents. Considerable development continues to take place in all categories of sensors, including research into the chemical and biological properties of seawater. Atmospheric temperature and humidity sensors must have shielding to protect them from heating by solar radiation, yet be open enough to ensure adequate ventilation. Pyroheliometers (radiometers) must remain clean and not become covered with dust, salt, or bird droppings. Wind anemometers must withstand exposure to salt spray. In the past, precipitation measurement presented formidable problems, but the tropical moorings across the Pacific have successfully measured rainfall. An acoustic method that measures the noise generated by raindrops as they hit the surface also shows promise. These new methods will help fill a serious gap in our knowledge of ocean climates. High winds and complex airflow patterns around ships have prevented the accurate measurement of precipitation over the oceans from commercial vessels.

Fig. 3. An autonomous profiler float, PALACE type. (*Webb Research Corp.*)

Underwater sensors traditionally measure temperature and velocity, also measuring pressure as an indication of depth. They must operate at high hydrostatic pressure and resist the corrosive effects of salt. These conditions continue to challenge ocean engineers, especially for multiyear deployments. Temperature sensors often sit inside the thermally conducting metal instrument housing. Velocity sensors come in two groups: directly measuring mechanical devices and acoustic sensors. The former come in many configurations. A common design consists of a rotor that spins around a vertical axis in response to the flow past it while a separate vane measures the direction of flow relative to the instrument and a compass inside monitors the orientation of the instrument with respect to magnetic north.

Acoustic sensors come in two types: local and profiling. Local sensors measure the difference in acoustic travel time between two points caused by the flow moving with the sound in one direction and opposite it in the other. With two or three such pairs, one can obtain the horizontal or all three components of motion. Profiling acoustic current meters depend upon the backscatter of signals from suspended matter in the water, typically zooplankton and small nekton. These instruments, called acoustic Doppler current profilers, send out short pulses in four oblique (relative to the vertical) beams and then immediately listen to the backscattered echoes. From the slight Doppler shifts of the return signals, such a sensor can determine the magnitude of flow in the direction of each beam, from which one can then calculate the corresponding horizontal and vertical motions. These instruments have proven very effective in studies of upper ocean currents by looking upward from subsurface moorings and downward from surface platforms. Drifters, floats, and profilers do not require current sensors since that information comes directly from their movement.

The widespread use of instrumentation has had the beneficial effect of focusing attention on the development of reliable and stable conductivity sensors for estimating salinity. While conductivity sensors used for moored buoys have existed for a long time, the need to ensure that they remain stable for years while operating over a wide range of pressures, as well as at the surface (where biological fouling remains a particular problem), has led to redoubled engineering efforts. These efforts have already led to considerable improvement. Temperature and salinity constitute the two most important indicators of water type. Changes in these properties along float trajectories provide valuable information on the mixing and exchange of waters between different regions. Progress in the development of other biologically important sensors for particulate matter, downwelling light from the surface, chlorophyll, dissolved oxygen and carbon dioxide, nutrients, and pH holds much promise for the future. *See* OCEAN CIRCULATION; OCEANOGRAPHIC VESSELS; REMOTE SENSING; SEAWATER; THERMISTOR. H. Thomas Rossby

Bibliography. H. O. Berteaux, *Coastal and Oceanic Buoy Engineering*, 1991; *Sea Technology*, monthly.

Insulation resistance testing

The measurement of direct current or voltage drop across an electrical insulating material to determine its resistance by the application of Ohm's law. The main purpose of electrical insulation is to isolate the energized components of a system from each other and from ground. This requires that the insulation resistance to the flow of current be as high as possible. Good insulators have resistances in the range of 10^6 to 10^{20} ohms, and must also have acceptable mechanical, chemical, and thermal properties. *See* ELECTRIC INSULATOR; ELECTRICAL INSULATION; ELECTRICAL RESISTANCE; OHM'S LAW.

Although insulating materials in practical use can be solids, liquids, or gases, resistance measurements generally refer to solid insulation. Insulating liquids, as well as gases, are tested by applying an electric field by means of two electrodes submerged in the fluid, and raising the voltage to determine its dielectric strength. The insulating properties of a liquid can also be stated in terms of its specific resistance (or resistivity in ohm-centimeters). *See* DIELECTRIC MATERIALS; ELECTRICAL RESISTIVITY.

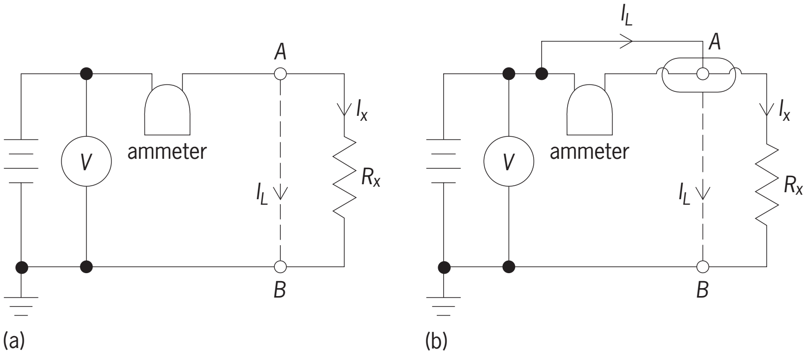


Fig. 1. Direct deflection method of resistance measurement (*a*) with no guard, and (*b*) with guard circuit applied at terminal *A* to eliminate the effects of leakage current.

Insulation resistance testing is performed either on samples of the insulation material or on the electrical equipment which uses the insulation. Tests on specimens are made to determine the material properties, while tests on the equipment are performed as a quality-control check at the factory or to determine the insulation condition while in service.

Insulation resistance in solids is made up of a combination of volume resistance and surface resistance. These two types of resistances arise because the current can physically flow through the volume of the insulation material or it can find other parallel paths over the surface of the insulation. Volume resistance can be used as an indirect measure of the quality of the material with regard to processing or to detect impurities. For example, indications of moisture content, degree of cure, mechanical continuity, or degradation can be obtained. Surface resistance measurements can be used to characterize the condition of the insulation surface, such as deposition of contaminants and insulation breakdown strength. The applicability of these measurements depends on the agreement of this type of data with actual field experience.

Factors affecting tests. In all cases, the conditions under which resistance measurements on insulators are performed must be fully specified. Several factors can significantly affect the test results. One of the most important of these factors is the geometry of the electrode system. With the correct selection of electrode and specimen dimensions, the surface and volume resistances can be measured separately. Temperature and humidity can change the resistance values by several orders of magnitude. To avoid this problem, the sample being tested can be conditioned prior to testing and the measurements can be made in the conditioning environment. A large error can also be introduced by stray surface currents. A guard system can be connected by forming a network of conductors which will shunt the stray currents away from the measuring circuit. The guard system is maintained at the same potential as that of the electrode to be guarded, but is insulated from it. Among the factors that may affect insulation resistance measurements are the nonuniformity of the material, the time period that the sample is energized, the magnitude and polarity of the voltage, the time required for charge to build up or decay in the sample, and the contour of the specimen.

Apparatus and methods. There is an extensive range of insulation test methods for both specimens and apparatus, and there are various accepted standards. Some typical tests are outlined below. Reproducible standard procedures should be used, and special care should be exercised in selecting the specimens and in their conditioning.

The measured value of insulation resistance will be most useful when the test samples and the electrodes have the same arrangement and are tested in a similar environment as that expected under

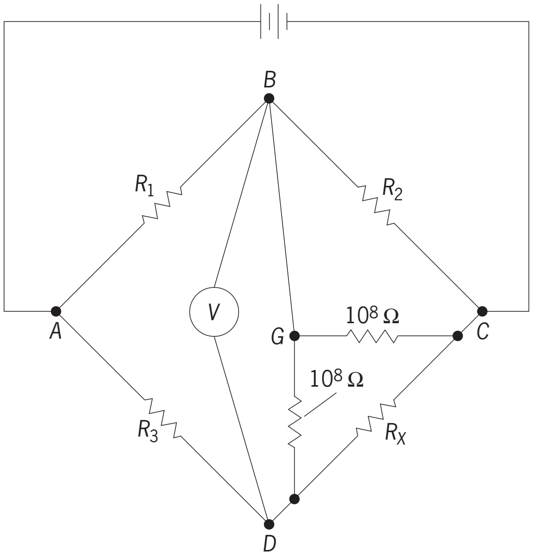


Fig. 2. High-resistance measurement with a guarded Wheatstone bridge.

actual use. A very stable direct current voltage source such as batteries is required. Instrumentation with a sensitivity of ±10% is generally considered adequate. Current-measuring devices commonly employed include galvanometers, direct-current (dc) amplifiers, and electrometers. Modern electronic instrumentation has simplified and improved insulation resistance measurements because of the high input impedance available with these devices, their fast response time, their memory storage capability, and their ease of operation. *See* ELECTROMETER; GALVANOMETER.

Direct deflection method. **Figure 1** shows a simplified version of the circuit used in this method with and without the application of a guarded terminal at *A*. The stray or leakage current is represented by I_L, and the insulation sample's resistance is R_X. The guard conductor surrounds terminal *A* and is connected to the battery side of the meter, so that no leakage current will flow through the meter, thus eliminating this source of error. Normally, the meter used is a sensitive galvanometer with a high-precision decade box (Ayrton shunt) in parallel to protect it and to provide several scale ranges. A switch can also be added to discharge a capacitive-type specimen after the test. The readings obtained are a measure of the volume resistance.

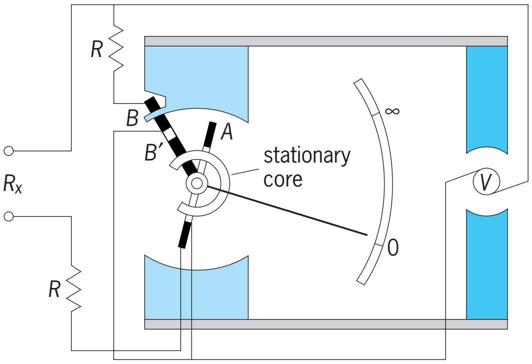


Fig. 3. Crank-driven Meggar circuit.

Loss-of-charge method. Insulation resistance is measured in this method by discharging a capacitor *C* through the sample to be measured. The charge *Q* remaining on the capacitor at the end of a specified time is measured with an electrometer or a galvanometer. Repeated measurements are necessary, making the procedure lengthy. Based on the voltage relationship $E = Q/C$, the value of resistance is found by plotting log *E* versus time. *See* CAPACITANCE; TIME CONSTANT.

Bridge method. A Wheatstone bridge circuit is utilized as shown in **Fig. 2**. The guard circuit in this case is made up by the two 10^8-ohm resistors connected to point *G*. The resistor from *G* to *D* shunts the detector, and the resistor from *G* to *C* shunts the *BC* arm, reducing stray current errors. *See* WHEATSTONE BRIDGE.

Meggar method. The Meggar is a megohmmeter that is widely used to measure the insulation resistance of rotating machinery, line insulators, bushings, and other equipment in the factory and on the field. Periodic measurements can provide valuable information to anticipate a serious insulation fault. The tests are made with an applied high voltage on the insulation of up to 5000 V direct current from a hand-driven generator or a rectified alternating-current (ac) supply. The Meggar consists of a permanent magnet crossed-coil meter (**Fig. 3**). A current coil *A* is connected in series with the unknown insulation resistance R_X. *B* and *B*′ are the potential coils connected to the generator and in series with the resistor *R*.

All coils are mounted on the pointer mechanism and move in the field set up by the generator's permanent magnets. The stationary ring core is mounted so that coils *B*′ and *A* rotate about it. When the generator is cranked, *B* and *B*′ will be energized. If *A* has no current through it, the pointer and coils mechanism will move to where the minimum flux will pass through them. This position is opposite the open portion or slot of the core, and the pointer will indicate infinity on the scale. If current flows in *A*, a torque will result and the three coils will move away from the slot of the core. Coils *B* and *B*′ produce opposing torques to this movement until a balance is reached. The balance point is determined by the current through *A*, which is directly dependent on the insulation resistance being tested. The reading is not dependent on the generator's voltage. *See* RESISTANCE MEASUREMENT.

T. C. Cheng

Bibliography. American Society for Testing and Materials, *Annual Book of ASTM Standards*, 1990; A. D. Helfrick and W. D. Cooper, *Modern Electronic Instrumentation and Measurement Techniques*, 1990; Institute of Electrical and Electronic Engineers, *Standard No. 118: Test Code for Resistance Measurements*, 1978; L. Schnell (ed.), *Technology of Electrical Measurements*, 1993; L. Thompson, *Electrical Measurements and Calibration*, 2d ed., 1994.

Insulin

One of the hormones produced by the beta cells of the pancreas, which are located in the islets of Langerhans.

Biosynthesis. Insulin participates in carbohydrate and fat metabolism. Along with insulinlike growth factors (IGF1 and IGF2), it regulates a variety of metabolic and growth-related effects in target tissues, including stimulation of glucose transport, glycogen synthesis, lipogenesis, gene transcription, and deoxyribonucleic acid (DNA) synthesis. Moreover, the insulin/IGF signaling system promotes cell growth and survival, and is essential for normal reproductive capacity. When insulin becomes unavailable or fails to function, diabetes ensues. *See* CARBOHYDRATE METABOLISM; HORMONE; LIPID METABOLISM; PANCREAS.

Human insulin is composed of 51 amino acids with a molecular mass of 6000. It is synthesized from the proinsulin gene located on human chromosome 11 (11p15.5). The proinsulin gene encodes a protein called pre-proinsulin, which is targeted to the endoplasmic reticulum where the amino-terminal signal peptide (referred to in **Fig. 1** as the signal sequence) is removed to produce proinsulin. Proinsulin displays only 10% of the biological activity of insulin. The A-chain and the B-chain that form insulin are produced by removal of the C-peptide (residues 33–63) from the middle of proinsulin (Fig. 1). The C-peptide facilitates proper folding of proinsulin so that three disulfide bonds form correctly between the A- and B-chains to stabilize the shape of insulin needed for biological activity. No biological function has been assigned to the C-peptide, although it is packaged and secreted with insulin in equimolar amounts. *See* GENE ACTION.

Physiological activity. The consumption of a carbohydrate meal causes an immediate rise in blood glucose concentration from 80 mg/dl to 130 mg/dl. Immediate insulin secretion then lowers blood glucose to the original level within 90 min. In healthy people, the blood contains about 0.2–1 ng/ml insulin before a meal and 1–5 ng/ml after the ingestion of a carbohydrate meal. Antibodies that specifically bind to insulin are used to determine precisely the concentration of insulin in blood. Similar methods measure the concentration of C-peptide in blood, which is used as an indicator that insulin is being produced in the beta cells, especially in people receiving insulin injections.

The blood glucose level is an important signal for insulin secretion. Glucose enters the pancreatic beta cells and increases the production of adenosine triphosphate (ATP), which directly stimulates insulin secretion. Other hormones released from cells in the small intestine during a meal travel to the beta cells to further promote insulin secretion. Upon secretion from the beta cells, insulin passes through the liver on its way to the peripheral tissues. Insulin-degrading enzyme is abundant in the liver, destroying about 50–60% of the secreted insulin as it passes through: the "half-life" of insulin in the body has been estimated to be 10–30 min. Insulin inhibits the production of glucose by the liver and increases the uptake of glucose by muscle and adipose tissues. Depending upon the tissue type, the glucose is used for energy (ATP production) or stored as glycogen or fat. *See* GLUCOSE.

Diabetes. The role of pancreatic secretions to control blood sugar levels was demonstrated in 1889. Insulin purified from the pancreas was first used to treat diabetic patients in 1921. Over the following decades, clinicians and scientists revealed the system-wide effects of insulin in liver, muscle, and adipose tissues. By the middle 1930s, the widespread use of insulin revealed important differences between insulin-dependent (type 1) and non-insulin-dependent (type 2) diabetes. Antibodies against insulin were used in the 1950s to confirm that hyperinsulinemia—elevated insulin in the blood—is characteristic of type 2 diabetes. Cell surface insulin receptors were first identified in the 1970s, and their tyrosine kinase activity was discovered 10 years later. Remarkably, this steady scientific progress has not stemmed the worldwide diabetes epidemic that will take a huge toll in premature morbidity and mortality in the twenty-first century.

The inability of beta cells to detect glucose or secrete enough insulin causes diabetes mellitus. Diabetes mellitus is a complex disorder that arises from various causes, including dysregulated glucose sensing or insulin secretion [maturity-onset diabetes of youth (MODY)], autoimmune-mediated β-cell destruction (type 1), or insufficient compensation for peripheral insulin resistance (type 2). MODY is a rare and usually mild form of diabetes caused by one of many well-defined genetic mutations. By contrast, the complete lack of insulin production that occurs in type 1 diabetes and in the late stages of type 2 diabetes is catastrophic. The most obvious problem is a 5- to 10-fold elevated blood glucose level, which causes life-threatening symptoms of diabetes mellitus—polydipsia (excessive thirst), polyphagia (excessive hunger), and polyuria (excessive urine production). The complete absence of insulin is accompanied by the breakdown of triglyceride into fatty acids. Fatty acids are useful fuel materials, but when present in large amounts they form beta-ketoacids in the liver, which acidifies the blood resulting in coma. In addition to an increased fat breakdown, an increase in the breakdown of tissue proteins occurs. Insulin injection reverses all of these acute problems by favoring storage and swift intake of glucose into the tissues, by decreasing the breakdown of stored fat, and by promoting protein synthesis. Optimized insulin treatment also reduces the occurrence of long-term complications that impair eye and kidney function and damages nerves and blood vessels.

Type 2 diabetes is the more prevalent form of the disease, affecting 18 million people in the United States alone. It develops more slowly and usually occurs in middle age, but is appearing with greater frequency in children and adolescents. Insulin resistance in muscle, liver, and adipose tissues is always associated with type 2 diabetes, and

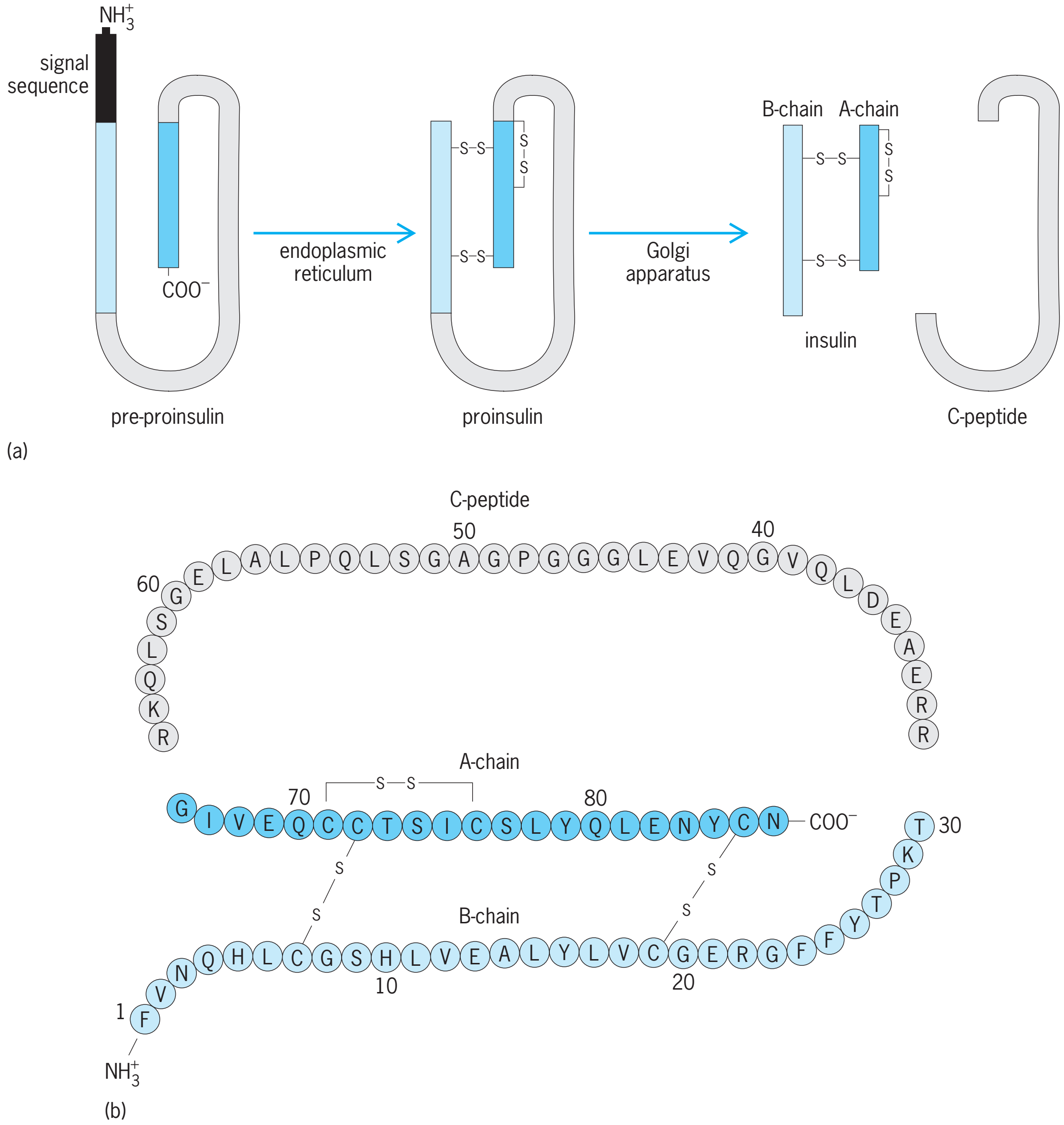


Fig. 1. Insulin biosynthesis. (*a*) Schematic diagram. Pre-proinsulin is synthesized from the insulin gene and directed into the endoplasmic reticulum by the signal sequence, where it folds into a proper conformation that is stabilized by three disulfide bonds (—S—S—). The signal sequence is removed, and proinsulin is further processed in the Golgi apparatus, where the C-peptide is removed and packaged with insulin for secretion. (*b*) Primary amino acid sequence of human insulin and the C-peptide.

hyperinsulinemia prevents gross decompensation of glucose homeostasis in the early stages of type 2 diabetes. However, insulin resistance is associated with a cohort of systemic disorders—dyslipidemia, hypertension, cardiovascular disease, female infertility and neurodegeneration—which usually develop before the disease is diagnosed. *See* DIABETES.

Commercial production. Human insulin was the first commercially available protein produced by recombinant DNA technology. The production is accomplished by inserting the gene for human insulin or proinsulin into a microorganism such as yeast or bacteria. While the microorganism grows, it produces the recombinant protein, which is purified and processed into biologically active insulin for injection. Treatment of diabetes with human insulin avoids the problems created by using insulin purified from animal sources, including the production of neutralizing antibodies.

The chemical characteristics of human insulin are exploited to produce insulin preparations that suit the treatment requirements of individual patients. Recombinant human insulin forms dimers in solution due to hydrogen bonding. Insulin dimers diffuse into the blood, where they separate into monomers that are biologically available 30–45 min after injection. The addition of zinc to insulin promotes the formation of insulin hexamers that require extra time to dissolve and form biologically active monomers: Insulin hexamers display a slower onset but a longer duration of activity (up to 24 h), resulting in less frequent need for insulin injections. The duration

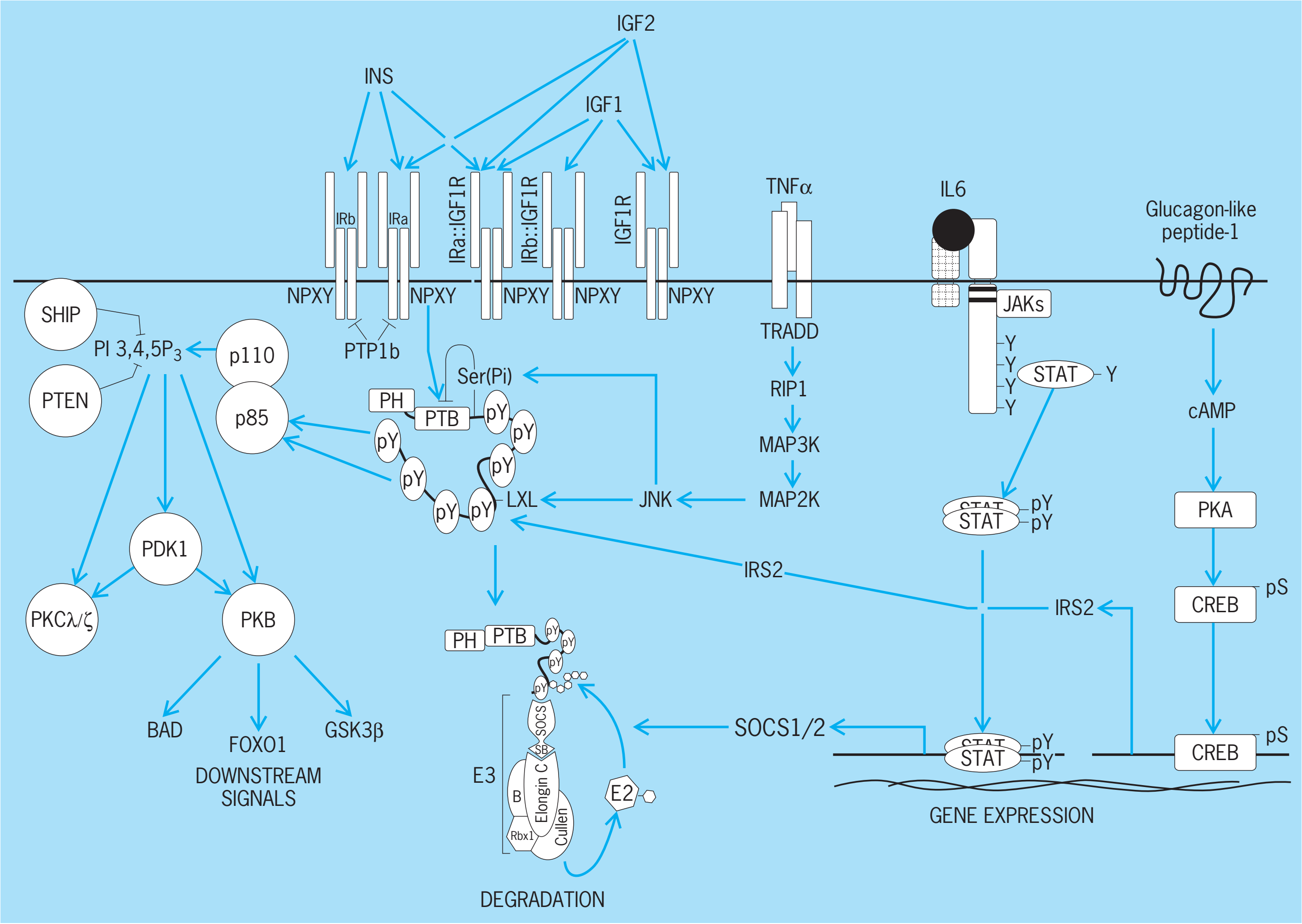


Fig. 2. Regulation of the insulin/IGF signaling pathway. The diagram shows the relation between the insulin receptor and its principal substrates, including some of their downstream elements and mechanisms of inhibition. The IRb predominates in adult tissues and binds insulin with highest affinity (<1 nM); IRa predominates during fetal life and in adult brain, binds insulin with lower affinity ($\sim$1 nM), but has a similar affinity for IGF2. The homologous IGF1 receptor binds IGF1/2 with high affinity (<1 nM). Hybrid receptors display intermediate binding properties: IRb::IGF1R binds IGF1 better than insulin, whereas IRa::IGF1R binds insulin and IGF1/2 with approximately equal affinity. The receptor β-subunits are tyrosine-phosphorylated during ligand binding and recruit substrates for phosphorylation, including IRS1/2 through an interaction between the phosphorylated NPXY motif on the receptor and the PTB domain. This interaction is inhibited by JNK-mediated phosphorylation of Ser^{312}, which causes insulin resistance. SOCS1/2 is an adapter protein that recruits IRS1/2 to an elongin B/C-based ubiquitin ligase, which promotes ubiquitination and degradation of the IRS-protein; proinflammatory cytokines like IL6 upregulate SOCS1/2. While proinflammatory cytokines inhibit IRS-protein signaling, cAMP agonists upregulate IRS2 through the activity of phosphorylated CREB. Tyrosine phosphorylation of IRS1/2 recruits and activates various SH2-domain-containing proteins, including the 85-kilodalton regulatory subunit of the PI 3-kinase. Activation of PI 3-kinase during association with IRS1/2 strongly stimulates the PKB cascades that phosphorylate various substrates, including BAD, GSK3β, and Foxo1.

of action of all forms of insulin is dependent on dose, site of injection, blood supply, temperature, and physical activity. Excess injected insulin causes severe hypoglycemia—blood sugar below its normal value—which damages neurons in the brain. For this reason, insulin should be injected subcutaneously and never intravenously or intramuscularly.

"Designer" insulin molecules with modified amino acid sequences can be produced by selectively changing the DNA sequence of the gene used in the process. These modifications alter the chemical behavior of insulin. For example, switching the order of the lysine and proline residues at the C-terminal end of the B-chain minimizes the formation of dimers and hexamers (Fig. 1). This designer insulin, called lyspro, displays almost immediate biological activity with short duration. It may be injected during a meal to prevent an excessive increase in plasma glucose after eating in certain patient populations. *See* GENETIC ENGINEERING.

Insulin receptors. Insulin action is initiated in cells by the binding of insulin to its cell receptor (**Fig. 2**). The insulin receptor is the prototype for a family of homologous integral membrane proteins composed of an extracellular insulin-binding domain that controls the activity of an intracellular tyrosine kinase. The 150-kilobase gene on chromosome 19, composed of 22 exons, encodes the human receptor precursor. During translation, two pro-receptors form disulfide-linked dimers that are glycosylated and cleaved to form a heterotetramer of two extracellular α-subunits and two transmembrane β-subunits.

Insulin binds to the juxtaposed α-subunits, facilitating ATP binding and tyrosine autophosphorylation in the β-subunit, which activates the kinase and recruits cellular substrates: the first steps of signal transduction.

Selective insulin binding is complicated by the production of two types of insulin receptors (IRa and IRb) by tissue-specific alternative splicing of exon 11. The type b insulin receptor that lacks exon 11 is mainly expressed in adult peripheral tissues (most tissues except the brain), where it selectively binds insulin with high affinity. The type a isoform that includes exon 11 is expressed in fetal tissues and the adult brain, where it binds insulin or insulinlike growth factor-2 (IGF2). Dysregulated splicing alters fetal growth patterns and contributes to rare forms of insulin resistance in adults. Homologous IGF1 receptors selectively bind IGF1 or IGF2, and are essential during development for about 50% of brain and body growth. However, hybrid receptors composed of an $\alpha\beta$-dimer of the IGF1 receptors and the IRb selectively bind IGF1. By contrast, hybrids between the IGF1R and the IRa bind all three ligands with similar affinities (Fig. 2).

Signaling. Most if not all insulin signals are produced or modulated through tyrosine phosphorylation of IRS1 or IRS2 (insulin receptor substrate 1 or 2). IRS-proteins are not enzymes but are composed of multiple interaction domains and phosphorylation motifs that generate regulatory signals. All IRS-proteins contain an NH_2-terminal pleckstrin homology (PH) domain (a domain of about 100 residues involved in intracellular signaling) adjacent to a phosphotyrosine-binding (PTB) domain, followed by a variable-length COOH-terminal tail with numerous tyrosine and serine phosphorylation sites. The PH and PTB domains couple the IRS-proteins to the activated insulin receptors. Tyrosine phosphorylation sites in the COOH-terminal part of IRS1/2 bind common effector proteins, including enzymes (PI 3-kinase, SHP2, and Fyn) or adapters (Grb-2, nck, crk, Shb, and others). Through these generic signaling proteins, IRS1/2 controls gene expression needed for carbohydrate and lipid metabolism, cell growth and survival, and the function of specialized neuroendocrine tissues that promote nutrient homeostasis and female fertility.

Tyrosine phosphorylation of IRS1/2 leads to the stimulation of the phosphatidylinositol 3-kinase (PI-3), which is composed of p85 and p110 subunits. PI-3 generates phosphatidylinositol 3,4,5-trisphosphate (PI $3,4,5P_3$) in the plasma membrane, which stimulates the activity of certain serine kinases in cells. One of these serine kinases, called protein kinase B (PKB), is especially important because it promotes the phosphorylation of other proteins, including BAD (important for cell survival), GSK3β (cell growth and glycogen synthesis), and Foxo1 (regulation of gene expression) [Fig. 2].

Resistance. Since insulin resistance is an underlying cause of metabolic dysregulation and diabetes, understanding its molecular basis is an important goal. Genetic mutations are obvious sources of lifelong insulin resistance, but they are difficult to identify and usually associated with rare metabolic disorders. The idea that inflammation is associated with insulin resistance has been known for a long time, which provides a framework to understand how diet, acute or chronic stress, and obesity might cause insulin resistance. Familiar proinflammatory cytokines, including interleukin-6 (IL6) and tumor necrosis factor alpha (TNFα), ordinarily secreted from leukocytes during inflammation or from adipocytes in obese people, promote insulin resistance. These cytokines stimulate serine phosphorylation of IRS1/2, which inhibit their ability to undergo tyrosine phosphorylation. These cytokines also promote ubiquitination of IRS1/2, which targets them for degradation. Insulin resistance is also caused by protein or lipid phosphatases, including PTP1B, SHIP2, or PTEN. PTP1B is especially interesting because it resides in the endoplasmic reticulum, where it dephosphorylates the insulin receptor. This mechanism of action appears to confer selectivity toward the insulin receptor, so inhibitors of PTP1B might be least likely to cause unregulated cell growth.

Peripheral insulin resistance alone rarely causes diabetes, as β-cell failure is an essential feature of the disease. Why β-cells frequently fail to compensate for insulin resistance was difficult to determine until the IRS2-branch of the insulin/IGF signaling system was found to be essential for β-cell growth, function, and survival. Transgenic mouse experiments clearly show that upregulation of IRS2 potentates compensatory insulin secretion in obese mice and prevents β-cell destruction. Recent work shows that the IRS2 gene in mice is strongly upregulated by cAMP agonists, including glucose and glucagon-like peptide-1 in β-cells, through pathways that activate CREB (Fig. 2). While many cAMP-mediated pathways oppose the action of insulin, the upregulation of IRS2 reveals an unexpected intersection of these important signals that prepares cells to respond during the next round of insulin secretion (Fig. 2). IRS2 upregulation might play a role in the beneficial health effects of many drugs that increase cellular cAMP levels. The positive effects of glucagon-like peptide-1 on β-cell function, peripheral insulin action, and neuronal function might be mediated in part by upregulation of IRS2.

Diabetes is a serious illness, but only one of the consequences of insulin resistance. The clinical effects of insulin resistance can be expected to increase, as the people become heavier and less fit. Whether better management of inflammatory responses can attenuate insulin resistance and diminish its consequences is an important area of investigation. Understanding how the IRS-proteins negotiate the conflicting signals generated during insulin and cytokine action might be a valuable starting point. However, too much insulin action might shorten our lives, so future work must better resolve the network of insulin responses that are generated in various tissues, and pinpoint the ones that prolong health as well as the ones that diminish it. Morris F. White

Bibliography. M. Bliss, *The Discovery of Insulin*, University of Chicago Press, 2003; M. F. White, Insulin signaling in health and disease, *Science*, 302:1710–1711, 2003; P. Zimmet, K. G. Alberti and J. Shaw, Global and societal implications of the diabetes epidemic, *Nature*, 414:782–787, 2001.

Intaglio (gemology)

The name given to the type of carved gemstone in which the figure is engraved into the surface of the stone, rather than left in relief by cutting away the background, as in a cameo. Intaglios are almost as old as recorded history, for this type of carving was popular in ancient Egypt in the form of cylinders. The cylinder, as well as the more familiar form of intaglio, was popular for impressing seals on sealing wax. Intaglios have been carved in a variety of gem materials, including emerald, crystalline quartz, hematite, and the various forms of chalcedony. *See* CAMEO; GEM.

Richard T. Liddicoat, Jr.

Integral equation

An equation of the form typified by Eq. (1). The major

$$\int_a^b K\left(x, y, \phi(y)\right) dy + f(x) = a(x)\phi(x) \qquad (1)$$

problem is to decide when there is a function $\Phi(x)$ which is a solution to the equation. Equations such as (1) arise from the analysis of ordinary differential equations. For example, it is easy to verify that the function $\Phi(t)$ is a solution of differential equation (2) if and only if Φ is a solution of integral equation (3).

$$\frac{d\phi}{dt} = F(t, \phi) \qquad \phi(0) = A \qquad (2)$$

$$\phi(t) = A + \int_0^t F(s, \phi(s))ds \qquad (3)$$

This equation is the same type as Eq. (1), with the function K given by Eq. (4).

$$K(t, s, \phi) = \begin{cases} F(s, \phi) & 0 \le s \le t \\ 0 & t < s \end{cases} \qquad (4)$$

Integral equation (1) is an equation of the first kind if $a(x) \equiv 1$, and an equation of the second kind if $a(x) \equiv 0$.

The domain of integration, which in Eq. (1) is the interval $[a, b]$, can be a domain in two or more variables, a curve, or even a surface. The unknown function Φ and the kernel K can be vector-valued functions, in which case Eq. (1) is called a system of integral equations.

Volterra equations. If $K(x, y, \Phi) = 0$ for $x \leqslant y \leqslant b$, Eq. (1) is called a Volterra equation. Thus Eq. (3) is an example of a Volterra equation. Under mild assumptions about the smoothness of $K(x, y, \Phi)$, for example, if K is continuous and if $|K(x, y, \Phi) - K(x, y, \Psi)| \leqslant A|\Phi - \Psi|$, where A is a constant, for all x, y in the domain of integration and for all Φ and Ψ—then the Volterra equation of the second kind, Eq. (5), can be

$$\phi(x) = \int_a^x K\left(x, y, \phi(y)\right) dy + f(x) \qquad (5)$$

solved by the method of successive approximations (also known as the method of Picard). The idea is to define a sequence of functions $\Phi_0, \Phi_1, \ldots$ according to the scheme of Eqs. (6). Under the assumed

$$\begin{aligned} \Phi_0(x) &= f(x) \\ \Phi_1(x) &= \int_0^x K\left(x, y, \Phi_0(y)\right) dy = f(x) \\ &\cdots \\ \Phi_{n+1}(x) &= \int_0^x K\left(x, y, \Phi_n(y)\right) dy + f(x) \end{aligned} \qquad (6)$$

conditions, it can be shown that this sequence converges to a function $\Phi(x)$ which is the only solution to Eq. (5). The method of Picard is fundamental to the study of integral equations, but occurs also in the study of many other functional equations.

Linear integral equations. A special case of some interest is that of linear integral equations. The function $K(x, y, \Phi)$ is a linear function of Φ. The linear equations of the first and second kind are shown in Eqs. (7) and (8) respectively. The function

$$f(x) = \int_a^b K(x, y)\phi(x)dy \qquad (7)$$

$$\phi(x) = f(x) + \lambda \int_a^b K(x, y)\phi(y)dy \qquad (8)$$

$K(x, y)$ is called the kernel, and the complex number λ is called the parameter. The equation of the second kind, Eq. (8), is homogeneous if $f(x) \equiv 0$. In typical cases, the homogeneous equations will have only the trivial solution $\Phi(x) \equiv 0$. For some values of the parameter λ, however, there will be nontrivial solutions. Such a value of the parameter λ is called a characteristic value for K, and the corresponding function Φ is called a characteristic function for K. These concepts are related to corresponding concepts in linear algebra and operator theory. The operator L_K acting on the function Φ is defined by Eq. (9). Then L_K is an example of a linear operator,

$$L_K\phi(x) = \int_a^b K(x, y)\phi(y)dy \qquad (9)$$

that is, if Φ and Ψ are functions and α and β are complex constants, then $L_K(\alpha\Phi + \beta\Psi) = \alpha L_K\Phi + \beta L_K\Psi$. If the function Φ is a characteristic function for K with corresponding characteristic value λ, then the homogeneous equation can be rewritten as Eq. (10).

$$L_K\phi = \frac{1}{\lambda}\phi \qquad (10)$$

In terms of linear operator theory, Eq. (10) means that $1/\lambda$ is an eigenvalue (proper value) for the linear operator L_K, and Φ is a corresponding eigenvector (proper vector), or in this case, eigenfunction.

A linear Volterra equation has no characteristic values. *See* EIGENFUNCTION; MATRIX CALCULUS; MATRIX THEORY; OPERATOR THEORY.

Fredholm equations. The integral equation, Eq. (7) or (8), is called a Fredholm equation, and $K(x, y)$ is called a Fredholm kernel if $\|K\| < \infty$, where $\|K\|$, the norm of K, is defined by Eq. (11). This is certainly

$$\|K\|^2 = \int_a^b \int_a^b |K(x, y)|^2 \, dx \, dy \qquad (11)$$

true if K is continuous or even bounded. Even some infinite discontinuities are allowable: for example, if $|K(x,y)| \leq M|x - y|^{-\alpha}$, where M is a constant and $\alpha < {}^1\!/_2$, then K is a Fredholm kernel.

Fredholm equations of the second kind for which the parameter satisfies the inequality $|\lambda| \, \|K\| < 1$ can be shown to have unique solutions by the method of Picard. The result is Eq. (12) where the resolvent $R_K(x, y; \lambda)$ is given by the Neumann series, Eqs. (13) and (14). Equation (12) can be rewritten in terms of the operator L_K as Eq.(15), when $L^2{}_K f = L_K(L_K f)$, $\ldots, L^{n+1}{}_K(f) = L_K(L^n{}_K f)$.

$$\phi(x) = f(x) + \lambda \int R_K(x, y; \lambda) f(y) dy \qquad (12)$$

$$\begin{gathered} R_K(x, y; \lambda) = \sum_{n=1}^{\infty} \lambda^{n-1} K^{(n)}(x, y) \\ K^1(x, y) = K(x, y) \end{gathered} \qquad (13)$$

$$K^{(n+1)}(x, y) = \int_a^b K(x, z) K^{(n)}(z, y) \, dz \qquad (14)$$

$$\phi = \sum_{n=0}^{\infty} \lambda^n L_K^n f \qquad (15)$$

As a result there are no characteristic values for which $|\lambda| \leqslant 1/\|K\|$. In general, for any constant $M > 0$ there are only finitely many characteristic values which satisfy $|\lambda| \leqslant M$. Further, only finitely many linearly independent characteristic functions are associated with each characteristic value.

For the kernel K the conjugate kernel is defined to be $K^*(x, y) = \overline{K(y, x)}$, where the bar denotes complex conjugation. It can be shown that λ is a characteristic value for the kernel K if and only if $\overline{\lambda}$ is a characteristic function for the conjugate kernel K^*. Furthermore, the number of linearly independent corresponding characteristic functions is the same for λ and K as for $\overline{\lambda}$ and K^*. Finally, inhomogeneous equation (16) has a solution if and only if f satisfies Eq. (17) for every solution ψ of conjugate homogeneous equation (18).

$$\phi(x) = f(x) + \lambda \int_a^b K(x, y) \phi(y) dy \qquad (16)$$

$$\int_a^b f(x) \overline{\psi(x)} \, dx = 0 \qquad (17)$$

$$\psi(x) = \lambda \int_a^b K^*(x, y) \psi(y) dy \qquad (18)$$

These results are a statement of the Fredholm alternative which may be stated more succinctly as follows: For a given value of λ, either λ is a characteristic value and the homogeneous equation has a nontrivial solution, or λ is not a characteristic value and the inhomogeneous equation has a solution for every choice of the function f. *See* DIFFERENTIAL EQUATION; INTEGRAL TRANSFORM; INTEGRATION. John Polking

Bibliography. C. Corduneanu, *Integral Equations and Applications*, 1991; J. Kondo, *Integral Equations*, 1992; A. C. Pipkin, *A Course on Integral Equations*, 1991; D. Porter and D. S. Stirling, *Integral Equations: A Practical Treatment, from Spectral Theory to Applications*, 1991.

Integral transform

An integral relation between two classes of functions. For example, a relation such as Eq. (1) is said to define an integral transform. More generally, the integral may be a multiple integral, and the functions f, G, ϕ may depend on a larger number of variables. Equation (1) is thought of as transforming a whole class, or space, of functions $\phi(y)$ into another class of functions $f(x)$. The function $G(x, y)$ is the kernel of the transform. One of the important uses of such a transform is based on the fact that a problem posed in one of the two spaces in question may be more easily solved in the other. For example, a differential equation to be solved for the function $\phi(y)$ may become an algebraic equation for the unknown function $f(x)$. *See* LAPLACE TRANSFORM.

$$f(x) = \int_{-\infty}^{\infty} G(x, y) \phi(y) \, dy \qquad (1)$$

The two basic problems for any integral transform are inversion and representation. In inversion the aim is to recover $\phi(y)$ from $f(x)$, the kernel $G(x,y)$ being known. That is, Eq. (1) is thought of as an integral equation (of the first kind) to be solved for the unknown function $\phi(y)$. A means of calculating $\phi(y)$ from $f(x)$ is called an inversion formula, and in its presence the transform achieves maximum utility, since explicit passage from each space to the other is thus assured. In representation the question is which functions $f(x)$ may be written or represented in the form (1). That is, one asks which functions $f(x)$ will make Eq. (1) solvable for $\phi(y)$. Usually this problem becomes more tractable when the solutions $\phi(y)$ are restricted to some subspace such as the class of positive or bounded functions.

Inversion theory. In many cases where the inversion problem has been solved, an inversion operator, such as a differential or integral operator, has been found, depending perhaps on a parameter, which accomplishes the inversion. It may be denoted by $O_t[f(x)]$. This means that for each value of the parameter t some operation O on $f(x)$ is defined. For example, t might be 3 and the operation might consist of computing the third derivative of $f(x)$. Suppose that O_t applied to Eq. (1) produces Eq. (2), where

$$O_t[f(x)] = \int_{-\infty}^{\infty} G_t(x, y)\phi(y)\,dy \tag{2}$$

$G_t(x, y)$ is a new kernel for each t. In favorable cases the integral (2) approaches $\phi(x)$ as t approaches a suitable limit, for example, when G_t approaches the Dirac δ-function $\delta(x - y)$.

An important special case of Eq. (1) is the convolution transform, when the kernel is a function of $(x - y)$, Eq. (3). An equivalent form of Eq. (3) is Eq. (4), since the change of variable $x = e^t$, $y = e^{-u}$

$$f(x) = \int_{-\infty}^{\infty} G(x - y)\phi(y)\,dy \tag{3}$$

$$F(x) = \int_{0}^{\infty} K(xy)\Phi(y)\,dy \tag{4}$$

carries Eq. (4) into Eq. (3) after a suitable change in notation.

Named transforms. Many of the classical transforms have been named for the original or principal investigator. A few of the more important ones can be listed with the name ordinarily attached to them.

For Eq. (1):

A. Bilateral Laplace	$G(x, y) = e^{-xy}$	
B. Fourier	$G(x, y) = e^{ixy}$	
C. Mellin	$G(x, y) = y^{x-1}$	$y > 0$
	$= 0$	$y < 0$
D. Stieltjes	$G(xy) = (x + 1)^{-1}$	$y > 0$
	$= 0$	$y < 0$

For Eq. (3):

E. Dirichlet	$G(x) = x^{-1} \sin x$	
F. Fejér	$G(x) = x^{-2}(\sin x)^2$	
G. Fractional (Weyl form	$G(x) = (-x)^{\mu-1}$	$x < 0$
	$= 0$	$x > 0$
H. Hilbert	$G(x) = (x)^{-1}$	
I. Picard	$G(x) = e^{-\lvert x\rvert}$	
J. Poisson	$G(x) = (1 + x^2)^{-1}$	
K. Laplace (unilateral)	$G(x) = e^x e^{-e^x}$	
L. Stieltjes	$G(x) = \operatorname{sech}(x/2)$	
M. Weierstrass or Gauss	$G(x) = e^{-x^2}$	

For Eq. (4):

N. Fourier cosine	$K(x) = \cos x$
O. Fourier sine	$K(x) = \sin x$
P. Hankel	$K(x) = \sqrt{x}J_v(x)$
Q. Laplace (unilateral	$K(x) = e^{-x}$
R. Meijer	$K(x) = \sqrt{x}K_v(x)$

Here $J_v(x)$ is a Bessel function and $K_v(x)$ a modified Bessel function. Certain transforms are listed twice because of their frequent appearance in two equivalent forms. In fact, all of the above are convolution transforms except A, B, and C. Also A and C are the same transform (set $y = e^{-t}$ in C).

Convolution inversion. The inversion theory for a large class of convolution transforms has been completed. Operational calculus is a guide to the method. If the symbol D stands for differentiation with respect to x but is nonetheless treated as a number in the familiar series represented by Eq. (5), one is led to make the definition in Eq. (6), since the series (6)

$$e^{tD} = \sum_{k=0}^{\infty} \frac{t^k}{k!} D^k \tag{5}$$

$$e^{tD} f(x) = \sum_{k=0}^{\infty} \frac{t^k}{k!} D^k f(x) = f(x + t) \tag{6}$$

is precisely the Maclaurin expansion of $f(x + t)$ if D^K is allowed to mean a kth derivative. *See* OPERATOR THEORY.

Now consider the bilateral Laplace transform of $G(t)$ in Eq. (1), denoting it by $1/E(s)$, as shown in Eq. (7). Replacing s by D and using Eq. (6), one

$$\frac{1}{E(s)} = \int_{-\infty}^{\infty} e^{-st} G(t)\,dt \tag{7}$$

has Eq. (8). If $E(D)$ were a number, one could solve Eq. (8) to obtain Eq. (9). Finally, reverting to the original meaning of D as a derivative, one has in Eq. (9) an inversion of Eq. (1) by means of a differential operator $E(D)$.

$$\begin{aligned} \frac{1}{E(D)}\phi(x) &= \int_{-\infty}^{\infty} e^{-tD}\phi(x)G(t)\,dt \\ &= \int_{-\infty}^{\infty} \phi(x - t)G(t)\,dt \\ &= \int_{-\infty}^{\infty} G(x - y)\phi(y)\,dy \\ &= f(x) \qquad x - t = y \end{aligned} \tag{8}$$

$$\phi(x) = E(D)f(x) \tag{9}$$

This argument is meant to be exploratory only, but the result is accurate for a large class of kernels G and their corresponding inversion functions E. In summary, the inversion function is the reciprocal of the bilateral Laplace transform of the kernel. It has been shown that the result is correct if, for example, $E(s)$ is the infinite product in Eq. (10), where $c \geqq 0$

$$E(s) = e^{bs - cs^2} \prod_{k=1}^{\infty} \left(1 - \frac{s}{a_k}\right) e^{s/a_k} \tag{10}$$

and the series of real constants

$$\sum_{k=1}^{\infty} a_k$$

converges.

For example, if $K(x) = e^{-x}$, then Eq. (4) is the Laplace transform. Expressed as a convolution transform as in Eq. (3), it becomes Eq. (11), where G

$$e^x F(e^x) = \int_{-\infty}^{\infty} G(x - y)\Phi(e^{-y})\,dy \tag{11}$$

is given in the above list as entry K. The bilateral Laplace transform of this kernel is the familiar gamma function, Eq. (12), whose reciprocal has

$$\Gamma(1 - s) = \int_{-\infty}^{\infty} e^{-st} G(t) dt = \int_{0}^{\infty} e^{-t} t^{-s}\,dt \tag{12}$$

a well-known expansion in the form of Eq. (10). In Eq. (13) γ is Euler's constant. In the present

$$E(s) = \frac{1}{\Gamma(1-s)} = e^{-\gamma s} \prod_{k=1}^{\infty} \left(1 - \frac{s}{k}\right) e^{s/k} \qquad (13)$$

example Eq. (9) becomes Eq. (14), or if $e^x = t$, Eq. (15) may be written. This familiar inversion for-

$$e^{-\gamma D} \prod_{k=1}^{\infty} \left(1 - \frac{D}{k}\right) e^{D/k} e^x F(e^x) = \Phi(e^{-x}) \qquad (14)$$

$$\lim_{k\to\infty} \frac{(-1)^k}{k!} F^{(k)} \left(\frac{k}{t}\right) \left(\frac{k}{t}\right)^{k+1} = \Phi(t) \qquad (15)$$

mula also serves to illustrate the operator O_t appearing in Eq. (2). In the present case the operator is a differential one, and the parameter t is an integer k which tends to ∞. *See* CONFORMAL MAPPING; INTEGRATION.

David V. Widder

Bibliography. M. I. Akhiezer, *Lectures on Integral Transforms*, 1992; L. C. Andrews and B. K. Shivamoggi, *Integral Transforms for Engineers and Applied Mathematicians*, 1987; M. Ya. Antimirov, A. A. Kolyshkin, and R. Vaillancourt, *Applied Integral Transforms*, 1993; A. J. Jerri, *Integral and Discrete Transforms with Applications and Error Analysis*, 1992.

Integrated-circuit filter

An electronic filter implemented as an integrated circuit, as contrasted with filters made by interconnecting discrete electrical components. The design of an integrated-circuit filter (also called simply an integrated filter) is constrained by the unavailability of certain types of components, such as piezoelectric resonators, that are often valuable in filtering. However, integrated filters can benefit from small size, close integration with other parts of a system, and the low cost of manufacturing very complex integrated circuits.

Mathematically, there are two broad types of filters: Continuous-time filters implement systems of linear ordinary differential equations, while discrete-time filters implement systems of difference equations. Both types can be integrated. *See* DIFFERENCE EQUATION; DIFFERENTIAL EQUATION.

Filters have many applications. A particularly important one is to smooth signal waveforms sufficiently to allow accurate sampling or to interpolate smoothly between given samples of a signal. Since the analog-to-digital converters that sample signals are usually made as integrated circuits (chips), it is often convenient to put the associated filters on the same chip. *See* ANALOG-TO-DIGITAL CONVERTER.

Passive filters, made by interconnecting inductors and capacitors, are not easily integrated because integrated-circuit inductors are usually of poor quality. They are compromised by being restricted to two-dimensional construction, by the resistance of the very thin metal available, and by proximity to a conductive substrate in which currents can be induced. These problems are less damaging at frequencies above 1 GHz, so microwave filters can be passive. This technology is most common when using gallium arsenide (GaAs) substrates, because higher speeds, insulating substrates, and high-conductivity metal are available there. These circuits are known as monolithic microwave integrated circuits (MMICs), where monolithic is a synonym for integrated. *See* MICROWAVE FILTER.

Active filters. Because amplifiers are very cheap in integrated-circuit technology, active filters are widely implemented. The five main types of active filter—active-RC, MOSFET-C, transconductance-C, switched-capacitor (or switched-C), and active-RLC—are distinguished by their frequency-sensitive components. The switched-capacitor filters operate on samples of signals, while the other types operate without sampling (in continuous time). There is also a trend toward digital filters, which are easily integrated but require that analog signals be converted to digital form, which in turn requires filtering. *See* DIGITAL FILTER.

The filter examples discussed below are second order, meaning that they implement a system of two linear differential or difference equations, or equivalently that their transfer functions can be written as ratios of polynomials of order two (quadratics). They are called biquadratic circuits because their transfer functions can be so written. Practical filters are often of higher order, but a filter of arbitrary order can be made by cascading (placing one after another) second-order circuits.

Active-RC filters. Discrete active-RC filters were widely used in the 1970s, and modern integrated filters are derived from them. The frequency-sensitive mechanism in active-RC filters is the charging of a capacitor C through a resistor R, giving a characteristic frequency $\omega_0 = 1/RC$ radians per second, at which the impedances of the resistor and capacitor are equal. Unfortunately, integrated-circuit manufacturing techniques do not control the product RC at all accurately, with variations of 20–50% being possible. This limits active-RC filtering to those applications where accuracy is unimportant, where external passive components are tolerable, or where tuning circuitry is available.

MOSFET-C filters. These filters replace the resistors of an active-RC filter with metal-oxide-semiconductor (MOS) transistors, in which a conducting channel along the surface can be enhanced or depleted by applying an electrical field from a gate electrode, thereby changing the resistance of the channel. The result is a tunable variant of an active-RC filter. A simple example (**Fig. 1**) consists of a loop of two integrators (V_2 is the integral of V_1, for example) implemented by amplifiers, which may be labeled A1 and A2, and their associated devices. An additional amplifier, A3, inverts the sign of the output of A2 to set loop gain correctly. Replacing all transistors with resistors (and ignoring the tuning control V_c) reduces this circuit to the Tow-Thomas biquad. *See* TRANSISTOR.

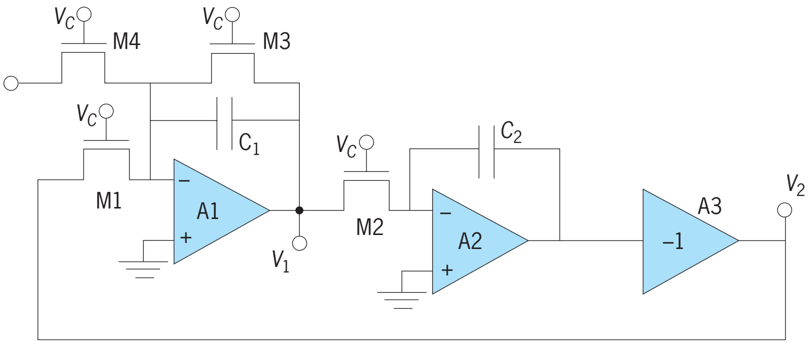


Fig. 1. MOSFET-C filter, consisting of a loop of two integrators.

If two of the transistors, M1 and M2, are equal in size, and each consists of Q parallel copies of the two other transistors in the circuit, M3 and M4, then, for a given control voltage V_c, M1 and M2 will simulate resistors of some value R and M3 and M4 will simulate resistors of value QR. Then when the two capacitors in the circuit, C_1 and C_2, are designed with $C_1 = C_2 = C$, the output of amplifier A1 will be a band-pass-filtered version of the input, with a center frequency $\omega_0 = 1/RC$, at which the input is passed unattenuated, and a bandwidth ω_0/Q between the frequencies at which the output voltage gain is $1/\sqrt{2}$. Band-pass filters are used to detect tones, while the output of amplifier A2 is a low-pass filter of the kind used to smooth signals before sampling. The control voltage V_c is generated with a separate circuit that adjusts the behavior of a reference MOSFET-C circuit with a simple known input.

MOSFETs are nonlinear resistors in that their resistances change significantly for voltage inputs of a few hundred millivolts. This behavior is partially controlled by using a differential or balanced version of the circuit discussed above, in which all signals are carried in both positive and negative polarities; this type of circuit tends to cancel quadratic nonlinearities, and also partially cancels interference from noise on the power supply and external sources. It also makes the amplifier A3 redundant, since differential signals can be inverted just by exchanging two wires.

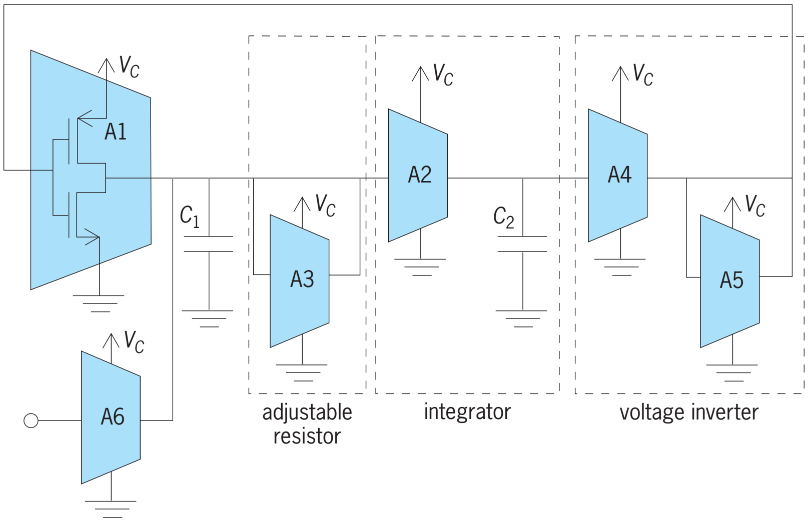


Fig. 2. Transconductance-C filter with amplifier A1 shown in detail.

Transconductance-C filters. These filters combine the functions of the amplifier and the simulated resistor into a transconductance amplifier, which is an amplifier whose output current (rather than output voltage) is proportional to its input voltage. The effect of M2 and A2 in the MOSFET-C filter (Fig. 1) is to force a current V_1/R into the feedback capacitor on A2; a transconductance amplifier with gain $1/R$ and input voltage V_1 has an output current V_1/R, so if it is connected to a capacitance C the same effect will be obtained. A simple transconductance-C filter (**Fig. 2**) can be made entirely from transconductance amplifiers (which may be labeled A1 to A6) and capacitors. Amplifiers that have their inputs and outputs connected together (A3 and A5) just act as grounded resistors, since they have a single signal terminal and it conducts a current proportional to the applied voltage. Since transconductors are usually adjustable, so is this simulated resistor. When one of these simulated resistors loads another transconductor (as A5 loads A4), an inverting amplifier is formed.

There are many circuits that approximate an ideal transconductance amplifier. One of the simplest (A1 in Fig. 2) consists of a complementary pair of *n*-channel and *p*-channel MOS transistors. Other, more practical implementations have a balanced form in which every signal is represented by a pair of equal and opposite voltages; just as for the MOSFET-C case, this reduces the effect of interference and even-order nonlinearities while eliminating the need for voltage inverter formed by A4 and A5. *See* AMPLIFIER; DIFFERENTIAL AMPLIFIER.

Transconductance amplifiers can be very simple and hence are capable of high-frequency operation (up to approximately 1 GHz) but tend to have poor linearity when designed for high speeds. The complementary MOS pair (Fig. 2) is one of the simplest and fastest but is difficult to adjust because the power-supply voltage acts as the control signal.

The transconductance-C filter can also be described in terms of gyrators, which are circuit-theoretic elements that simulate inductors at one port when loaded by capacitors at another (Fig. 2). In this view, A1, A2, A4, and A5 are a gyrator loaded by C_2, and the result is that the capacitor C_1 appears to be in parallel with a simulated inductor and a resistor (A3). *See* GYRATOR.

A technique known as active-RLC filtering combines the ideas of active filtering with the use of physical inductors (made as spirals of metallization on the top layer of the chip). In this method, amplifiers, connected to simulate negative resistors, are used to enhance the performance of the inductors, whose losses can be modeled (to a first approximation) as being caused by a parallel positive resistance.

Switched-capacitor filters. The basic Tow-Thomas loop can also be implemented in discrete time, that is, on samples of a signal (**Fig. 3**). Each of the

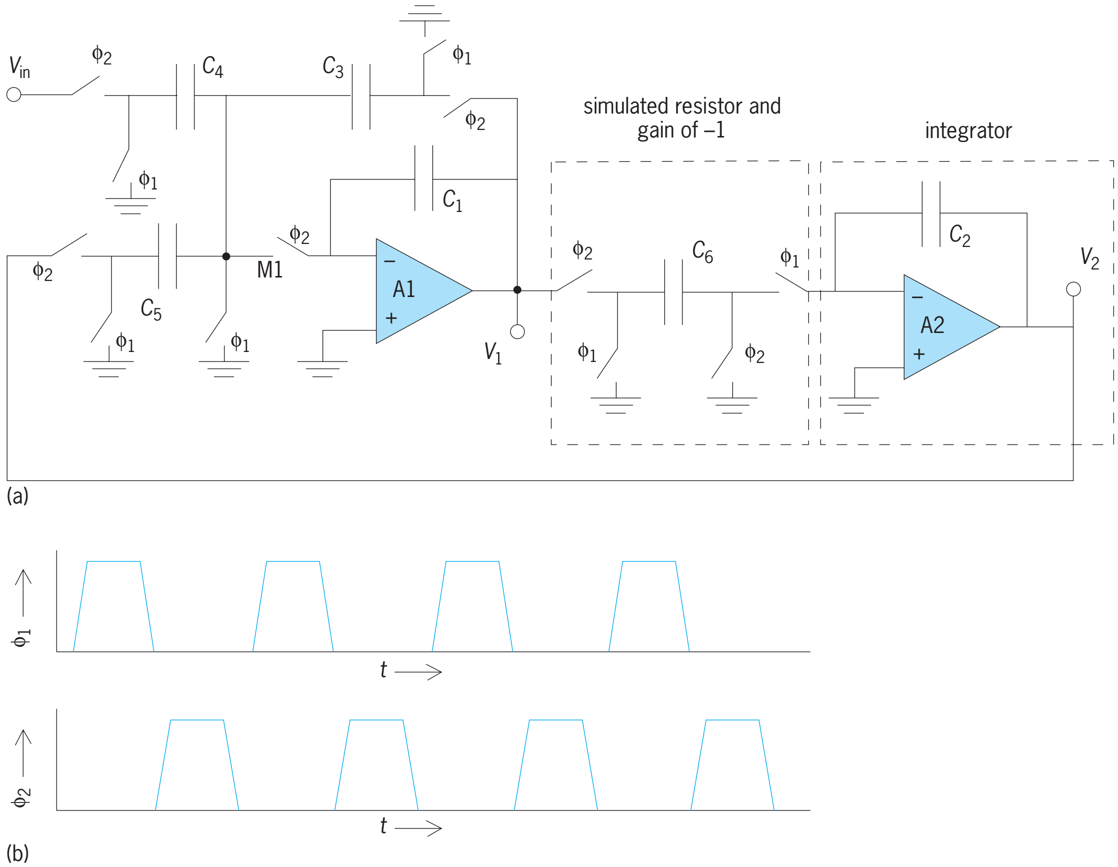


Fig. 3. Switched-capacitor filter. (*a*) Circuit diagram. (*b*) Nonoverlapping clock phases ϕ_1 and ϕ_2.

simulated resistors is now replaced with a capacitor and an arrangement of switches clocked on one or the other of two nonoverlapping phases, ϕ_1 and ϕ_2. A capacitor C_6, for example, is charged up by a voltage V_1 on each phase ϕ_2, then discharged into the input of an amplifier A2 on each phase ϕ_1. Transferring a packet of charge on every clock cycle is a way of making current flow, and the fact that the charge transferred on each cycle is proportional to voltage makes the switched capacitor C_6 behave like a resistor. Thus C_6 and its switches replace the transistor M2 in the MOSFET-C filter (Fig. 1). The switch arrangement on C_6 also inverts the capacitor between phases, thus getting the effect of the voltage inverter in the MOSFET-C filter without using an amplifier.

In a similar way, a capacitor C_5 and its switches replace the transistor M1 in the MOSFET-C filter, while two capacitors C_3 and C_4 replace the transistors M3 and M4, respectively. The switching arrangement at the input of A1 is slightly different, so that voltage inversion does not occur. A simple design technique applies the idea that a switched capacitor simulates a resistor of value $R = 1/(Cf_s)$, where f_s is the sampling (clock) frequency. Under this first-order approximation the center frequency of the filter is $\omega_0 = (C_5C_6/C_1C_2)f_s$. An exact design style is now preferred, in which the circuit is seen to implement a set of difference equations, rather than differential equations. The voltage on capacitor C_2, for example, changes by $(C_5/C_2)V_1$ from the end of one phase ϕ_2 to the end of the next. However, it is difficult to predict the precise value of the voltage V_2 at intermediate times, because there are large transient currents at the beginning of each clock phase that cause the operational amplifiers to behave nonideally. Thus switched-capacitor filters produce known outputs at discrete times (the end of phase ϕ_2, in this example), while the MOSFET-C and transconductance-C filters of (Figs. 1 and 2) provide well-defined outputs continuously (in continuous time).

The primary advantage of switched-capacitor filters is that they can be very accurate, since critical frequencies are determined by the product of a clock frequency and a ratio of capacitors (rather than a single capacitor). Switched-capacitor filters are probably the most prevalent integrated filters. Most telephone systems, for example, use them to smooth signals before sampling them for digital transmission.

A key problem for switched-capacitor filters is that the switches, implemented as MOSFETs, inject small amounts of charge into the circuit when they turn off. This creates constant offset errors, couples noise in from power supplies, and can produce nonlinear distortion. Just as for MOSFET-C and transconductance-C filters, the trend is to implement switched-capacitor filters with fully balanced circuits, in which these effects are partially cancelled. *See* SWITCHED CAPACITOR CIRCUIT.

Current-mode filters. In these filters, the signals are represented by current levels rather than voltage

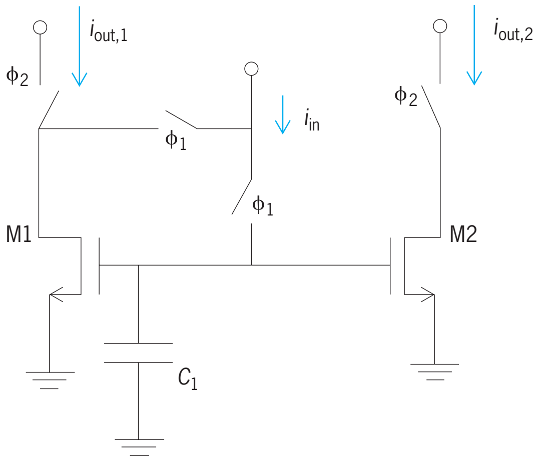


Fig. 4. Sampled-current circuit.

levels. A key building block of these filters is the current copier (**Fig. 4**). During phase ϕ_1, any difference between the input current i_{IN} and transistor drain current is fed back to correct the gate voltage of a transistor M1, so that a voltage is established on a parasitic capacitance C_1 that forces the transistor to operate at the input current. Then, during phase ϕ_2, the transistor is switched to drive this same current into the output. Placing a second transistor M2 in parallel with the copier device M1 allows for a second output that can be scaled by a constant factor (the ratio of device sizes) from the first; and signals represented by currents can be added just by connecting them together. Given the operations of scaling, addition, and copying, an arbitrary set of difference equations (and hence an arbitrary filter) can be implemented. The principal advantage of this technology over that of switched capacitors is that it can make use of the nonlinear gate-source parasitic capacitance that exists in any process, rather than needing a high-quality capacitor. A disadvantage comes from the lower accuracy of ratios set by transistor sizes. *See* ELECTRIC FILTER; INTEGRATED CIRCUITS. Martin Snelgrove

Bibliography. S. K. Mitra and C. F. Kurth (eds.), *Miniaturized and Integrated Filters*, 1989; R. Schaumann, M. S. Ghausi, and K. R. Laker, *Design of Analog Filters*, 1990; A. S. Sedra and K. C. Smith, *Microelectronic Circuits*, 4th ed., 1997.

Integrated circuits

Miniature electronic circuits produced within and upon a single semiconductor crystal, usually silicon. Integrated circuits (ICs) range in complexity from simple logic circuits and amplifiers, about 1.3 mm (0.05 in.) square, to very large scale integrated circuits (VLSI), up to about 25 mm (1 in.) square. They can contain over 10^9 transistors and other components that provide computer memory circuits and complex logic subsystems such as microcomputer central processor units (microprocessors) and digital signal processing (DSP). *See* SEMICONDUCTOR; SILICON.

Key inventions leading to the invention of the integrated circuit include the invention of the bipolar transistor by John Bardeen, Walter Brattain, and William Shockley in 1947; the invention of junction isolation by Kurt Lehovec; and the invention of silicon planar technology by Jean Horni.

Jack Kilby was the first to demonstrate an integrated circuit in 1958, using germanium and attaching small wires to connect transistors, diodes, and capacitors; however, it was the approach by Robert Noyce, which included patterned aluminum metal interconnections on silicon (**Fig. 1**), that made the integrated circuit revolution possible. Noyce may be considered the leader of this revolution, not only for his role in inventing the integrated circuit but also for his role in promoting integrated circuit development and applications.

Since the mid-1960s, integrated circuits have become the primary components of most electronic systems, and they are used in almost every piece of hardware that has some electronics. Their low cost, high reliability, and speed have been essential in furthering the wide use of digital applications. Microprocessors have spread the use of computer technology to personal computers, instruments, business machines, automobiles, consumer electronics, and other equipment. Other common uses of very large scale integrated circuits are in pocket calculators, electronic watches, cell phones, digital cameras, personal digital assistants (PDA's), appliance controls, and Internet circuits. For analog signal processing, integrated subsystems such as FM stereo demodulators and switched-capacitor filters are made. As improvements in the technology enable the development of more complex circuits, the challenge becomes not only one of how to build these circuits but also one of what to build, so that there is a demand for the new features and performance that will continue to drive the industry's growth. *See* CALCULATORS; DIGITAL COMPUTER; ELECTRONICS; MICROCOMPUTER; MICROPROCESSOR; MOBILE COMMUNICATIONS; SIGNAL PROCESSING; SWITCHED CAPACITOR CIRCUIT; WATCH.

Integrated circuits consist of the combination

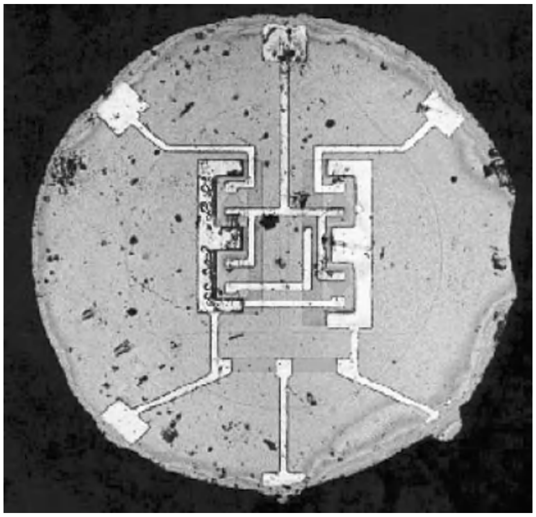


Fig. 1. Original integrated circuit, with aluminum interconnections on silicon. (*G. Moore, ISSCC '03, Intel Corp.*)

of active electronic devices such as transistors and diodes with passive components such as resistors and capacitors, within and upon a single semiconductor crystal. The construction of these elements within the semiconductor is achieved through the introduction of electrically active impurities into well-defined regions of the semiconductor. The fabrication of integrated circuits thus involves such processes as wet chemical etching, vapor-phase deposition of semiconductors and insulators, oxidation, ion implantation, vacuum deposition, sputtering, plasma etching, chemical mechanical polishing (CMP), rapid thermal processing (RTP), and lithography patterning techniques involving thin films of resist and optical exposure. *See* CRYSTAL GROWTH; DIFFUSION; ION IMPLANTATION; MICROLITHOGRAPHY; SEMICONDUCTOR HETEROSTRUCTURES; SPUTTERING.

Generally, integrated circuits are not straightforward replacements of electronic circuits assembled from discrete components. They represent an extension of the technology by which silicon planar transistors are made. Because of this, transistors or modifications of transistor structures are the primary devices of integrated circuits. Methods of fabricating good-quality resistors and capacitors have been devised, but the third major type of passive component, inductors, must be simulated with complex circuitry or added to the integrated circuit as discrete components. *See* TRANSISTOR.

Digital logic circuits were the easiest to adapt to these design changes. The first of these, such as inverters and gates, were produced in the early 1960s primarily for miniaturization of missile guidance computers and other aerospace systems. Analog circuits, often called linear integrated circuits, did not become commercially practical until several years later because of their heavy dependence on passive components such as resistors and capacitors. The first good-quality operational amplifiers for analog computers and instruments were produced in 1966. *See* ANALOG COMPUTER; LOGIC CIRCUITS; OPERATIONAL AMPLIFIER.

Types of Circuits

Silicon integrated circuits can be classified into two groups on the basis of the type of transistors that they employ: bipolar integrated circuits, in which the principal element is the bipolar junction transistor; and metal oxide semiconductor (MOS) integrated circuits, in which the principal element is the MOS transistor. Both depend upon the construction of a desired pattern of electrically active impurities within the semiconductor body, and upon the formation of an interconnection pattern of metal films on the surface of the semiconductor.

Bipolar circuits are generally used where highest logic speed is desired, and MOS for largest-scale integration or lowest power dissipation. Linear circuits have used bipolar technology extensively, but MOS devices are becoming much more important. High-performance bipolar transistors and complementary MOS (CMOS) transistors have been combined on the same chip (BiCMOS) to obtain circuits combining high speed and high density.

Integrated circuits are also made using other semiconductor materials, especially gallium arsenide (GaAs). These circuits tend to be more expensive but have much higher speeds. It is also much harder to integrate the large number of transistors on GaAs integrated circuits than can be done with silicon MOS. This article focuses on silicon MOS circuits, which probably account for over 95% of the integrated circuits currently produced (2006).

MOS integrated circuits. This major class of integrated circuits is called MOS because its principal device is a metal oxide semiconductor field-effect transistor (MOSFET). It is more suitable for very large scale integration than bipolar circuits, because MOS transistors are self-isolating and can have an average size of less than $6\lambda^3$, where λ is the minimal pattern that can be defined lithographically. This has made it practical to use over 10^9 transistors per chip. Because of this high-density capability, MOS transistors are used for high-density random-access memories (RAMs; **Fig. 2**), nonvolatile flash memories, and microprocessors. *See* COMPUTER STORAGE TECHNOLOGY; SEMICONDUCTOR MEMORIES.

Several major types of MOS device fabrication technologies have been developed since the mid-1960s: (1) metal-gate *p*-channel MOS (PMOS), which uses aluminum for electrodes and interconnections; (2) silicon-gate *p*-channel MOS, which employs polycrystalline silicon for gate electrodes and the first interconnection layer; (3) *n*-channel MOS (NMOS), which is usually silicon gate; and (4) complementary MOS (CMOS), which employs both *p*-channel and *n*-channel devices. The silicon gate CMOS is currently (2006) the dominant technology, and the term MOS usually refers to CMOS technology.

Both conceptually and structurally, the MOS transistor is a much simpler device than the bipolar transistor. In fact, its principle of operation has been known since the late 1930s, and the research effort

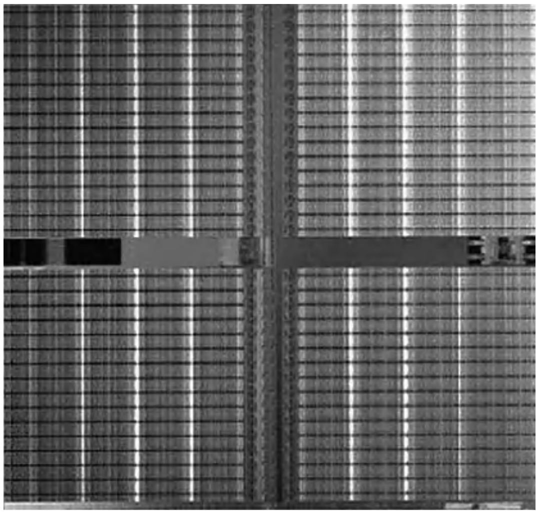


Fig. 2. Photomicrograph of metal oxide semiconductor (MOS) very large scale integrated circuit used as 153-megabit static random-access memory, using 45-nm technology (that is, with 45-nm feature sizes). Cell area = 0.346 μm^2, chip area = 119 mm^2. Chip has over 10^9 transistors. (*M. Bohr, Intel Corp.*)

that led to the discovery of the bipolar transistor was originally aimed at developing the MOS transistor. What kept this simple device from commercial utilization until 1964 is the fact that it depends on the properties of the semiconductor surface for its operation, while the bipolar transistor depends principally on the bulk properties of the semiconductor crystal. Hence MOS transistors became practical only when understanding and control of the properties of the oxidized silicon surface had been perfected to a very great degree.

CMOS technology. A simple CMOS inverter circuit is shown in **Fig. 3** and a circuit schematic in **Fig. 4**. The gates of the *n*-channel and *p*-channel transistors are connected together, as are the drains. The common gate connection is the input node, while the common drain connection is the output node. A capacitor is added to the output node to model the loading expected from the subsequent stages on typical circuits.

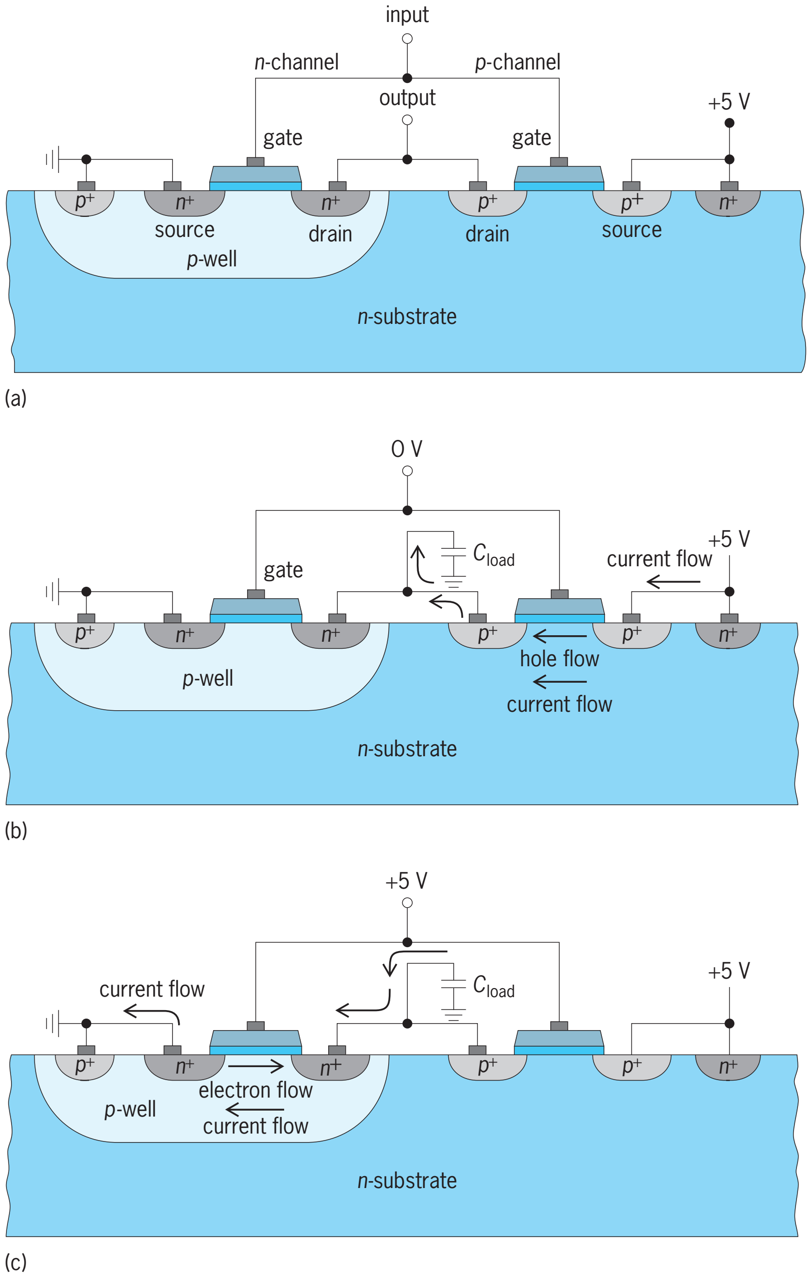


Fig. 3. Simple CMOS inverter circuit. (*a*) Schematic cross section. (*b*) Current flow when input is low at 0 V. (*c*) Current flow when input is high at 5 V.

When the input node is in the low state (0 V), the *n*-channel gate-to-source voltage is 0 V, while the *p*-channel gate to source voltage is −5 V. The *n*-channel transistor requires a positive gate-to-source voltage that is greater than the transistor threshold voltage (typically 0.5–1 V) before it will start conducting current between the drain and source. Thus, with a 0-V gate-to-source voltage it will be off, and no current will flow through the drain and source regions. The *p*-channel transistor, however, requires a negative voltage between the gate and source that is less than its threshold voltage (typically −0.5 to −1 V). When the input is put to the low state at 0 V, the −5-V gate-to-source potential is clearly less than the threshold voltage, and the *p*-channel will be turned on, conducting current from the source to the drain, and thereby charging up the loading capacitor. Once the capacitor is charged to the high state at 5 V, the transistor will no longer conduct because there will no longer be a potential difference between the source and drain regions.

When the input is now put to the high state at 5 V, just the opposite occurs. The *n*-channel transistor will be turned on, while the *p*-channel will be off. This will allow the load capacitor to discharge through the *n*-channel transistor, resulting in the output voltage dropping from a high state at 5 V to a low state at 0 V. Again, once there is no potential difference between the drain and source (capacitor discharged to 0 V), the current flow will stop, and the circuit will be stable.

This simple circuit illustrates a very important feature of CMOS circuits. Once the loading capacitor has been either charged to 5 V or discharged back to 0 V, there is no current flow since one of the transistors will be in the off state, and the standby power is very low. The important invention of CMOS digital logic was proposed as early as 1963 and became the dominant technology after 1980. The small stand-by power consumption is the reason for the high popularity of CMOS because of heat dissipation concerns. None of the other MOS technologies offers this feature without complex circuit techniques, and even then will typically not match the low standby power of CMOS. Bipolar circuits require even more power than these other MOS technologies. The price for the lower power of CMOS is the additional fabrication steps required (5–10% more) when compared to NMOS.

Integrated Optical Devices

Semiconductors have long been used as light sensors. Advances in very large scale integration have enabled large pixel arrays of light sensors, such as digital cameras and camcorders, to be made commercially. MOS is now preferred for sensor arrays because it permits large amounts of control logic to be fabricated in the same circuit as the sensors. *See* TELEVISION CAMERA; TELEVISION CAMERA TUBE.

In MOS optical arrays, there are two basic cell types: MOS devices or charge-coupled devices

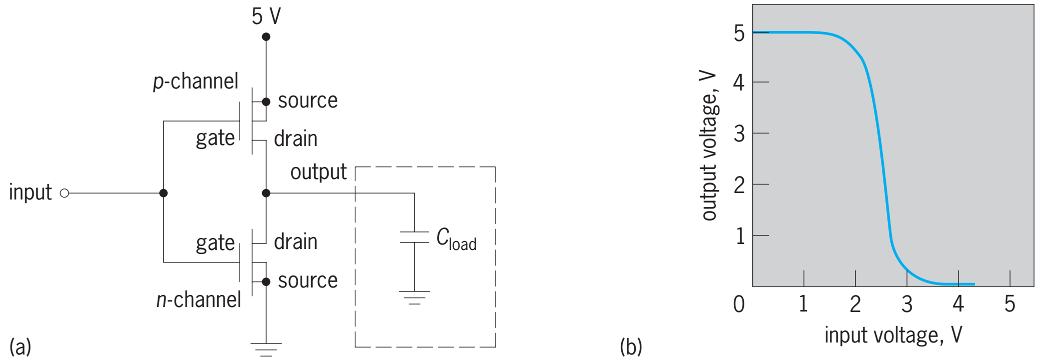


Fig. 4. **Characteristics of the CMOS inverter circuit of Fig. 3. (*a*) Circuit diagram. (*b*) Transfer characteristics.**

(CCDs). The sensor portion of the circuit views the scene through a transparent window. Light variations in the scene viewed cause variations in charge, producing signals transferable in shift-register fashion to processing stages such as amplifiers. These arrays have numerous applications, from measuring and sorting objects on production lines to optical character readers, which automatically translate printed and written information to digital computer input codes, such as the bar-code system and optical character recognition (OCR). *See* CHARACTER RECOGNITION; CHARGE-COUPLED DEVICES; COMPUTER VISION; INTEGRATED OPTICS; OPTICAL DETECTORS.

Microprocessors

The VLSI development having the most profound effect on electronic equipment economics and design in general is the microprocessor. The complexity and capability of these circuits, introduced in 1971, have expanded greatly. Initially these integrated circuits were applied to dedicated functions, with the tailoring of that function being done with the ROM circuits on the chip. However, there are now many different types of microprocessor circuits, ranging from the simple microcontroller types of functions to very complex computers on a chip such as the 64-bit microprocessor (**Fig. 5**).

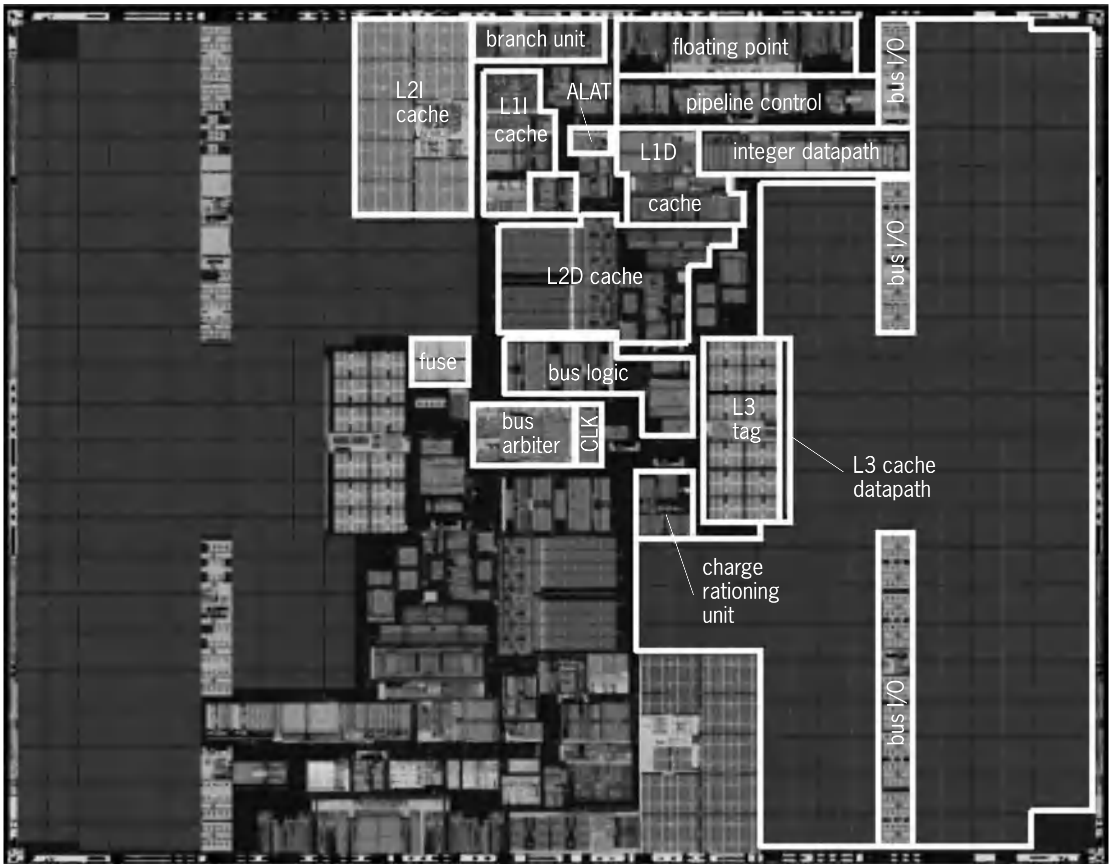


Fig. 5. **Photomicrograph of a 64-bit MOS microprocessor, the Dual Core Intel® Itanium® processor (codenamed Montecito), which uses 1.7 × 10^9 transistors. ALAT = advanced local address table, CLK = clock. (*Intel Corp.*)**

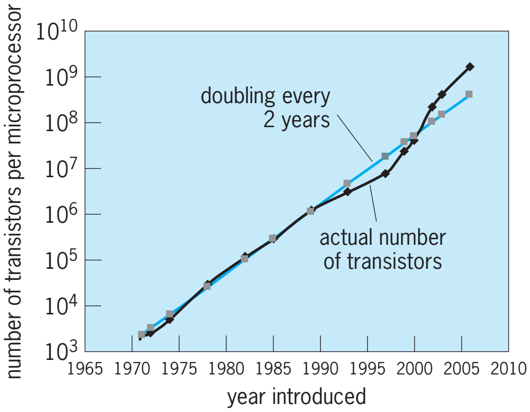


Fig. 6. Exponential growth in number of transistors for microprocessors. Actual number of transistors per microprocessor is compared with number that would have resulted from doubling every 2 years. (*Intel Corp.*)

This rapid development has been a result of both the technology which allows more circuits to be placed on a chip, and circuit innovation. The technology has increased the number of transistors on a chip from a few thousand in 1971 to over 10^9 in 2006, corresponding to a doubling every 2 years (**Fig. 6**). The circuit innovations include both digital techniques and the combination of analog and digital processing functions on the same chip. Both analog-to-digital and digital-to-analog data conversion circuits have been combined with the digital processor into one integrated circuit. The major digital innovations include more cache memory and parallel processing, with the example in Fig. 5 using two core processors.

Fabrication

Integrated-circuit fabrication begins with a thin, polished slice of high-purity, single-crystal semiconductor (usually silicon) and employs a combination of physical and chemical processes to create the integrated-circuit structures described above. Junctions are formed in the silicon slice by the processes of thermal diffusion and ion implantation. Electrical isolation between devices on the integrated circuit is achieved with insulating layers grown by thermal oxidation or deposited by chemical deposition. Conductor layers to provide the necessary electrical connections on the integrated circuit are obtained by a variety of deposition techniques. Precision lithographic processes are used throughout the fabrication sequence to define the geometric features required.

Requirements. The integrated circuit fabrication process is very sensitive to both particulate and impurity contamination. Airborne particulates must be minimized during the fabrication sequence, since even small ($\leq$0.05-micrometer) particles on the wafer surface can cause defects. A particulate-free fabrication ambient is normally achieved by the use of vertical laminar-flow clean rooms along with wafer transport systems that are clean, self-contained units that interlock with the equipment, reducing exposure to the environment. Lint-free garments are worn to minimize operator-borne particles (**Fig. 7**). To minimize impurity contamination effects, the chemicals, solvents, and metals that are used must be of the highest possible purity (electronic grade).

The precision and cleanliness requirements of integrated-circuit processing necessitate high discipline throughout the process sequence. This is achieved by extensive operator training, in-process tests and inspection with continual feedback, and a high degree of equipment calibration and control. The physical environment and operator attitude in an integrated-circuit fabrication facility are important factors for successful operation.

Simplified CMOS process flow. The principal steps in fabricating the simple CMOS inverter circuit of Fig. 3 are illustrated in **Fig. 8**, using an older (1980s) process flow. The process starts with the highly polished single slice of silicon crystal. For CMOS, the first step is typically to put in the *p*-well regions for the *n*-channel devices, if the original silicon wafer is *n*-type (Fig. 8*a*). It is also possible to start with a *p*-type wafer in which *n*-well regions are introduced for the *p*-channel transistors. While this discussion will be confined to the *p*-well case, most CMOS integrated circuits currently use wells for both the *n*-channel and *p*-channel transistors, with a *p*-type substrate starting material.

After the wells have been introduced, thick oxide layers (0.5–1.0 μm thick, called the field oxide) are grown in specific regions to isolate the transistor junctions from each other (Fig. 8*b*). Next the transistor gate oxides (typically 10–50 nanometer thick) are

Fig. 7. Person with lint-free garments in a vertical laminar-flow clean room for integrated-circuit fabrication with 300-mm (12-in.) wafer. (Personnel do not handle wafers in this manner. This was done just for the photograph.) (*Intel Corp.*)

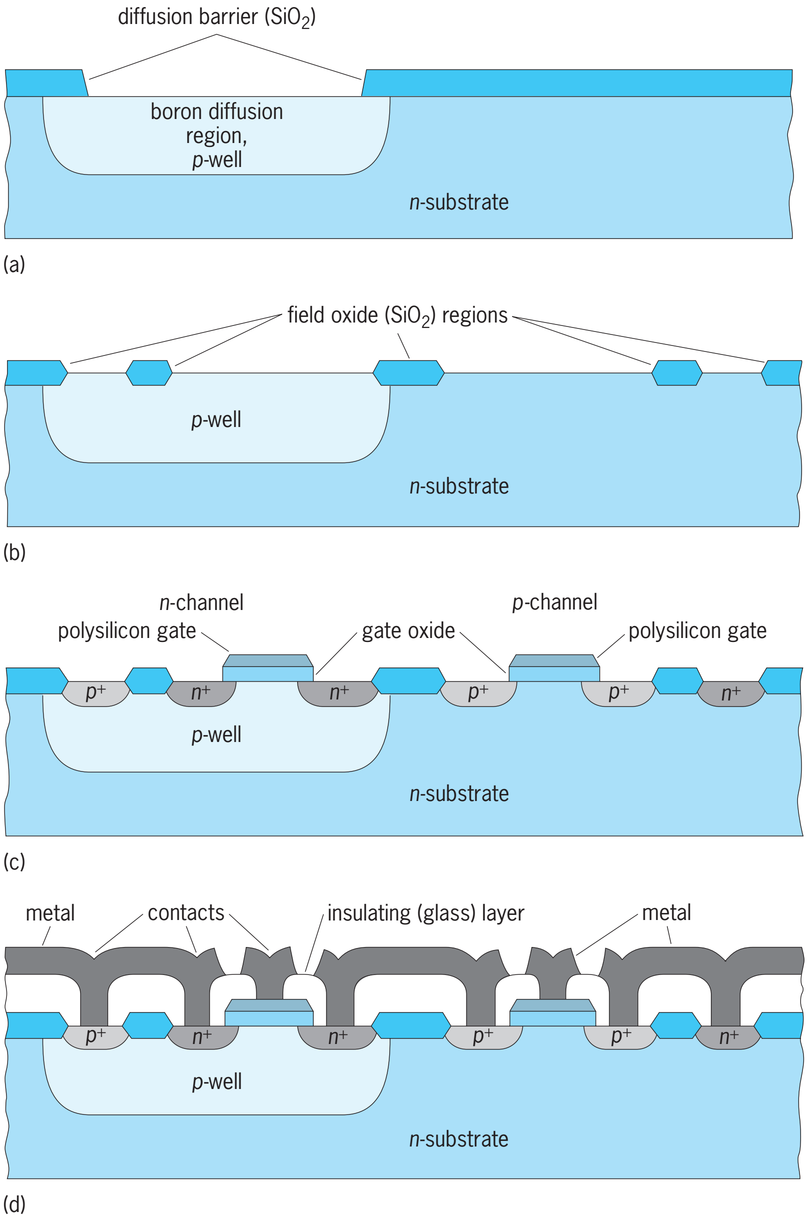


Fig. 8. Steps in fabrication of CMOS inverter circuit. (*a*) Introduction of *p* well. (*b*) Growth of field oxides. (*c*) Addition of gate oxides, polysilicon gates, and junctions. (*d*) Deposit of insulating layer, opening of contacts, and deposit of metal.

grown. Prior to the silicon gate deposition, implants may be used to control the threshold voltages of the transistors. Polysilicon is then deposited for the gate regions and doped with impurities to reduce the resistance. The gate regions are then defined by using photoresist and lithography, and the polysilicon and thin gate oxide is removed where it is not wanted. The transistors are now ready for the introduction of impurities for the source and drain junctions. With CMOS, separate masking steps are required to introduce the phosphorus or arsenic for the *n*-channel source and drain junctions, and the boron for the *p*-channel source and drain junctions (Fig. 8*c*).

An insulating layer (typically a glass) to isolate the polysilicon gate conductor from the aluminum metallization is now deposited. Openings in the glass layer are made for the metal to make contact to the polysilicon lines and source/drain junctions. The metal is then deposited, and the regions to be removed are etched away after using photoresist and lithography techniques (Fig. 8*d*). Finally a scratch-protection passivation layer is deposited over the metal, with openings made at the bond pad areas for the bond wire connections.

The isolation of the transistors is accomplished by reverse-biased junctions. The *n*-substrate (or *n*-well) is always at the highest chip voltage, while the *p*-well (or *p*-substrate) is always at the lowest chip voltage.

While the previous example is useful for developing the concepts of an integrated-circuit process flow, it does not match the current technology (**Fig. 9**), which includes about seven copper metal layers. The newer technologies use chemical-mechanical polishing (CMP) extensively throughout the process to keep the surface flat, starting with field oxide region. Chemical mechanical polishing is also used to remove the unwanted copper metal in the dual damascene process.

Trends and Process Changes

While the previous CMOS process flow is useful as an educational example to help highlight the fabrication process, the reality of current integrated-circuit technology is much more complex. The complexity of integrated circuits has been increasing exponentially since they were introduced in the early 1960s. This exponential growth was first recognized by Gordon Moore in 1965 and, as mentioned above, has been proceeding with a doubling about every 2 years, which has become known as Moore's law (**Fig. 10**).

Integrated-circuit fabrication technology changes rapidly due to the steadily decreasing feature size of the individual circuit elements, the improvements in yields allowing large chip sizes, the advances in equipment required for the finer features, and the increasing wafer sizes. Because of these rapid changes, a fabrication facility will remain state of the art in capability for no more than 2 years without making major equipment and process changes.

The size of the circuit elements has been decreasing as improvements in the lithography and etching techniques are incorporated (**Fig. 11***a*), amounting to a 30% reduction of the feature size every 2 years. This results in the area of the static RAM (SRAM) cell decreasing by about 50% every 2 years (Fig. 11*b*).

In addition to the increase in density, the wafer size has been increasing (**Fig. 12**), which makes it more economical to produce the large die sizes. Whereas in 1970 50-mm (2-in.) wafers were common, by 1980 100-mm (4-in.) wafers were the choice for new manufacturing lines. During the mid-1980s, the 150-mm (6-in.) wafer was the direction of new manufacturing lines, while in 1989 the 200-mm (8-in.) wafer began to be used. Starting about 2000, 300-mm (12-in.) wafers were being used in production. The larger wafer, plus the smaller feature sizes of the newer technologies, continues to decrease the cost of integrated circuits.

The economic drivers of the integrated circuit

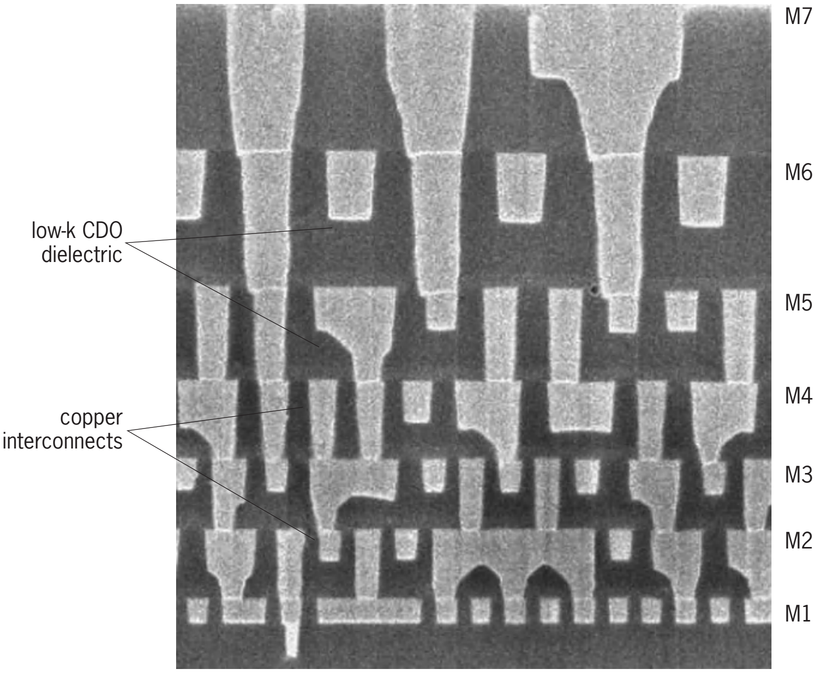


Fig. 9. Cross-sectional view using scanning electron microscope of chip manufactured by 90-nm process (that is, with 90-nm feature sizes), using seven copper metal layers. Low-permittivity (low-*k*) carbon-doped oxide (CDO) dielectric is used. (*G. Moore, ISSCC '03, Intel Corp.*)

industry have been cost and performance. As the technology has allowed faster and more complex functions on a chip, and feature size and yield have kept the cost from rising, the customer has been able to get more performance for the same or less cost. This is illustrated in **Figs. 13** and **14**, which show that while the cost of lithographic tools and the overall fabrication cost are rising the transistor cost and the performance cost, as measured in millions of instructions per second (MIPS), have been decreasing due to the smaller geometries and the efficiencies of scaling.

Historic process developments. These scaling trends could not have been achieved without major improvements in the technology. Along with improved lithography and transistor gate oxides, there were other major technology changes that made the scaling possible, and this trend must continue for the scaling to continue into the future.

The first major development in MOS integrated circuits was the understanding and development of quality gate oxides in the 1960s. This was from the work of many people, and resulted in an understanding of the silicon surface and the development of stable thermal oxides for the gate insulator. Initially, gate oxide thicknesses were greater than 100 nm (larger than the typical minimum feature size on today's technology), while currently (2006) they are of the order of 1–2 nm.

Another important development in the late 1960s was the use of polycrystalline silicon (polysilicon) for the metal gate. Previously, aluminum was used, which severely limited the processing temperatures that could be used after deposition. With polysilicon, self-aligned source-drain junctions could be fabricated, and polysilicon doping could better match threshold voltages for CMOS transistors. Polysilicon is still being used with a nickel salicide cap (discussed below).

Field oxides using the local oxidation process (LOCOS) with the self-align twin-well process were developed during the 1970s. This process uses a silicon nitride layer to prevent oxidation. This practice greatly lowered the surface step between the field oxide and the active region.

Ion implantation of elements was developed in the 1960s and 1970s, and became the primary method to introduce dopant impurities into the silicon in the 1980s. Both the dose and depth of the impurities can be controlled much better with implantation. It also leads naturally to source and drain tilt implants to

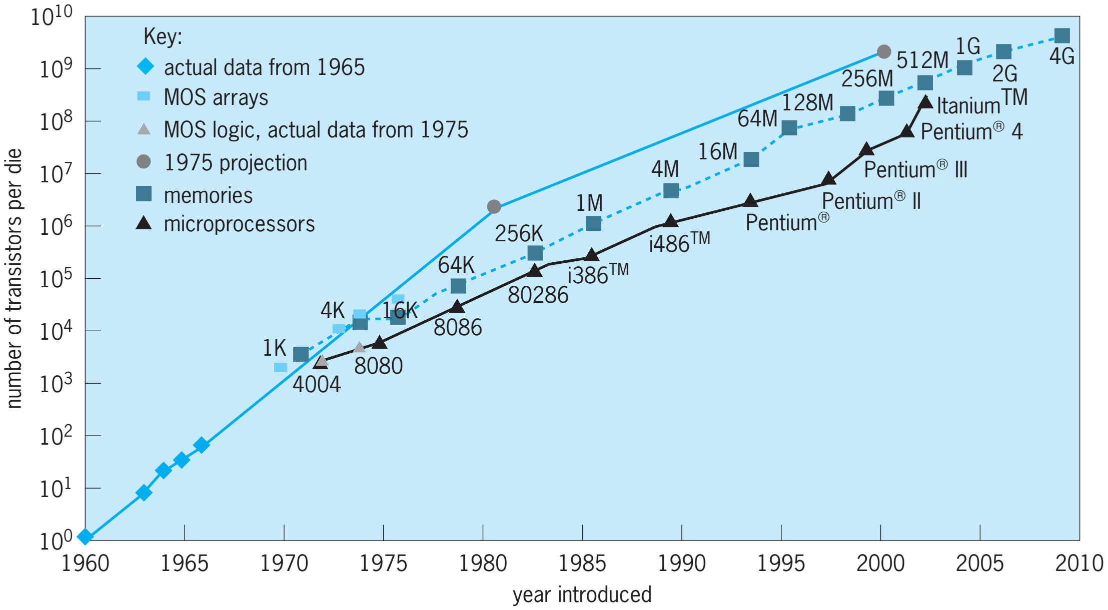


Fig. 10. Exponential growth in integrated circuit complexity, which is doubling about every 2 years. (*G. Moore, ISSCC '03, Intel Corp.*)

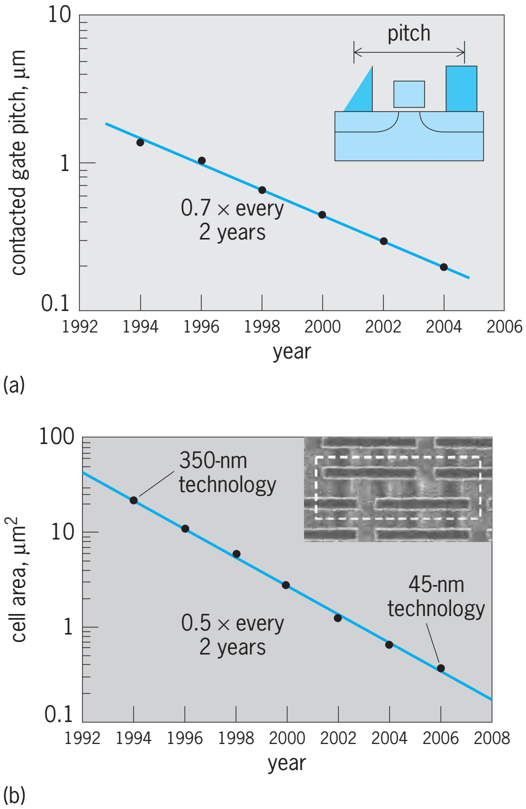


Fig. 11. Exponential trends of decreasing size and area. (*a*) Trend in contacted gate pitch (defined in the inset). Solid line indicates multiplication by factor of 0.7 every 2 years. (*b*) Trend in area of static random-access memory cell. Solid line indicates multiplication by factor of 0.5 every 2 years. Inset shows cell of 45-nm generation, introduced in 2006, with area of 0.346 μm, manufactured by dry-etching technique with ultraviolet light of 193-nm wavelength. (*P. Bai, Intel Corp.*)

control the punchthrough and short-channel effects (SCE) of submicrometer transistors.

Lithography changed significantly with the introduction of 5–10x step and repeat projection tools. Now one could make the mask 5–10 times larger and have better resolution and alignment control on the wafer. Instead of exposing the whole wafer at once, the exposure is into a small area on the wafer that can better be optimized, and this exposure is followed by stepping and repeating to the other areas on the wafer, until the total wafer has been exposed. Combining this technology with laser alignment, Ebeam mask generation, optical proximity correction, phase-shift masks, shorter wavelengths, and high-index immersion has allowed the printing of 22-nm feature sizes using optical techniques.

During the 1980s, increasing the die yields on the wafer became a major effort. Prior to this, the defect densities were greater that 1 defect/cm^2. Such a defect density meant that a die that was approximately 1 cm (0.4 in.) square could not be manufactured, since almost all the dice would then have a defect. The area of the microprocessor chip of Fig. 5 is almost 6 cm^2, which requires the defect density to be of the order of 0.1 defect/cm^2 or less for reasonable yields to be achieved. Thus, defect densities have decreased, at the same time that feature sizes have decreased and the complexity of the process has increased (the number of masking steps has more than doubled since 1980)—which is a remarkable achievement. This accomplishment required the introduction of new monitoring techniques during the processing of wafers (including new laser scanning systems), major modifications to processing equipment (to reduce defects introduced during processing), and new analysis techniques [such as fine-focused ion beam (FIB) techniques].

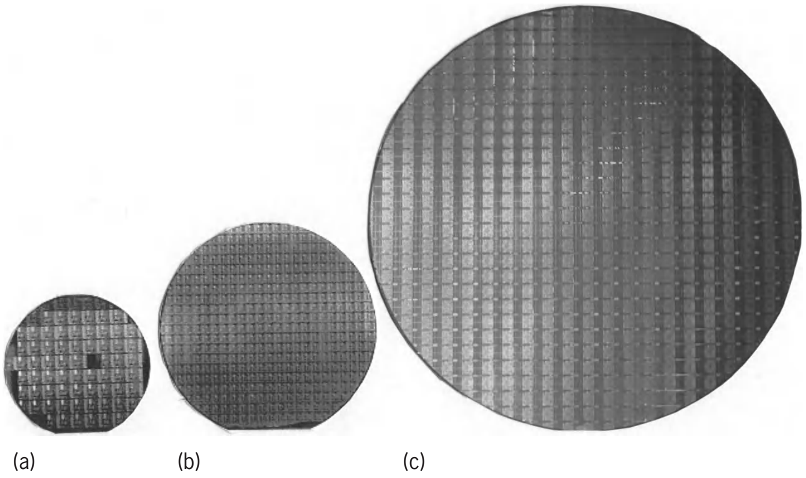


Fig. 12. Increase in wafer sizes, showing the increased number of dice (chips) per wafer available when increasing the wafer area. (*a*) 100-mm (4-in.) wafer. (*b*) 150-mm (6-in.) wafer. (c) 300-mm (12-in.) wafer. (*Intel Corp.*)

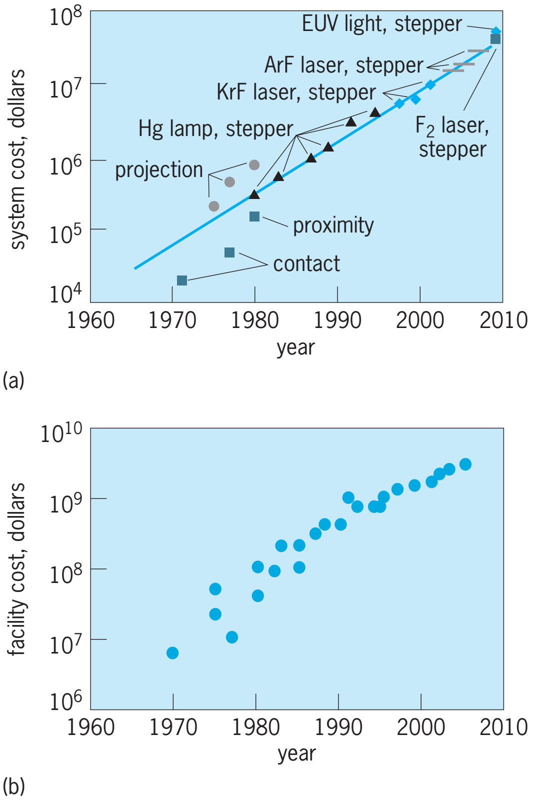


Fig. 13. Trends in (*a*) lithographic tool cost and (*b*) new fabrication costs in the integrated circuit industry. (*IC Knowledge LLC*)

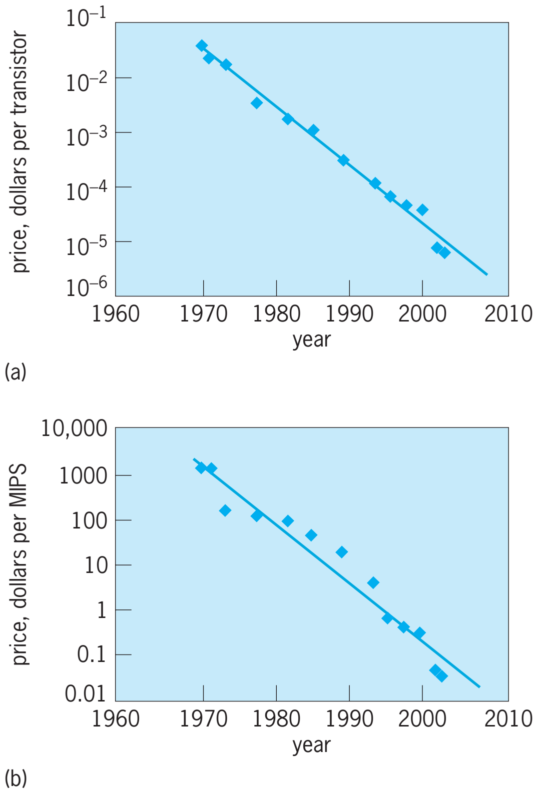


Fig. 14. Exponential trends in the price of the integrated circuit. (*a*) Price per transistor. (*b*) Price per MIPS (millions of instructions per second). (*S. Borkar, Intel Corp.*)

Dry-etching techniques using plasma became available in the 1980s and have become the primary etching techniques. In comparison with wet-etching techniques, their advantage lies in their ability to create large-aspect-ratio features without undercut. Without them, it would be impossible to pattern features smaller than 0.5 μm.

To reduce the polysilicon electrical resistance and improve the source/drain junction contact resistance, self-aligned silicide (salicide) formation on the polysilicon and junctions was developed during the 1980s. Initially, the salicide was formed using titanium deposition. However, currently nickel is the preferred metal. The silicide formation requires rapid thermal processing (RTP), which has been another important technology development and has been added to other parts of the process.

During the 1980s, a planarization technique involving chemical mechanical polishing (CMP) of the glass insulating layers between the metal layers was developed. This advance combined with tungsten plugs for the interconnection of the aluminum metal layers allowed the number of metal layers to increase from the two-layer practical limit at the time and dramatically reduced the lithographic and etching problems for metal layers.

During the 1990s, plasma etching into the silicon allowed for deeper field isolation without a large step on the surface, which is called the deep trench. This practice has since been combined with chemical mechanical polishing to remove the surface step altogether. The technique is also used in dynamic RAMs (DRAMs) to increase the area for the capacitors in the cells.

Because of its lower resistance and higher current capability, copper metal has always been preferred over aluminum metal. However, due to the limited number of good dry etch techniques for copper and the need for a diffusion-barrier layer. the dual damascene process had to be developed before copper could be practical. During this process, trenches are etched into the insulating dielectric layers between the metal layers (Fig. 9), which are first covered with a thin barrier seed layer, and then filled with copper using an electroplating process. The excess copper is then removed using chemical mechanical polishing techniques. Thus, copper can be removed without etching, which was always a major problem.

Prospective process developments. As the technology continues to shrink, the printing of these small features is becoming a major problem with the existing lithography tools. The wavelength of the light used in lithography has not decreased as significantly in recent technologies as previously (**Fig. 15**). Only with the special techniques of subresolution optical proximity correction (OPC), phase-shifted masks, and high-index immersion is it possible to print 45-nm features using 193-nm wavelength. Looking forward, it appears that extreme ultraviolet (EUV) radiation at a wavelength of 13 nm will be required at some point.

Another area of scaling concern is the transistor performance. As the gate oxide has been scaled to less that 2 nm, the gate current leakage has been increasing due to quantum tunneling effects. Also, surface effects increase the effective electrical thickness so that the capacitance coupling into the channel is less than classically predicted from physical oxide thickness, resulting in fewer free carriers, which means less channel current. To overcome these problems, high-*k* (high-permitivity) dielectrics are being investigated. In **Fig. 16**, the thicker experimental

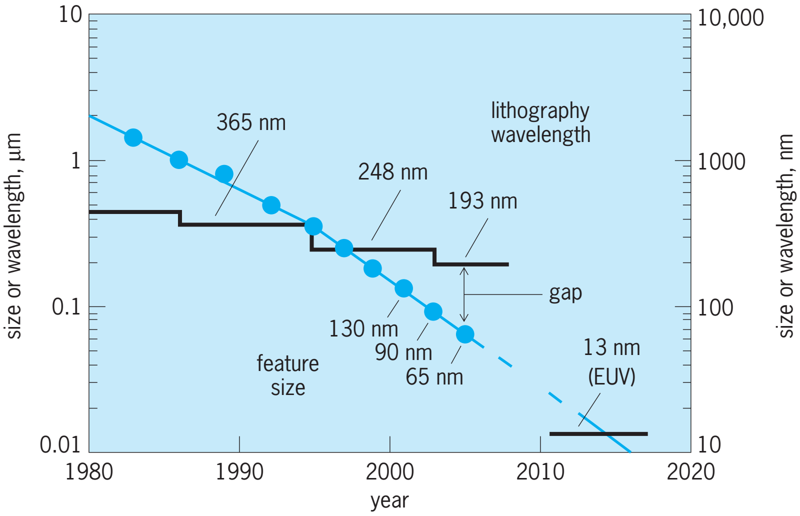


Fig. 15. Trends of feature size with lithography, compared with the wavelength of light used in lithography. (*P. Bai, Intel Corp.*)

high-*k* dielectric has more capacitance and over 100 times less leakage than the thinner silicon dioxide layer. *See* DIELECTRIC MATERIALS; PERMITTIVITY; TUNNELING IN SOLIDS.

Special efforts have been made to improve the carrier mobility of the transistors. As the carrier mobility increases, the current drive of the transistors increases, reducing delay, which increases the speed. The use of enhanced strain to improve mobility is illustrated in **Fig. 17** for both *n*-channel and *p*-channel transistors, where the NMOS strain is introduced through a tensile capping layer on the edges of the gate and PMOS strain is introduced using germanium-silicon (GeSi) in the channel region. Additional methods may need to be developed that continue to improve the mobility.

Another approach to improve the transistor performance is to use the additional sides of the transistor to carry current. An example is the tri-gate transistor (**Fig. 18**), whose drive currents are 50% higher than those of the planar transistor.

The delay in sending signals across the chip is related not just to the transistor drive but also to the load, which is mostly capacitance from the metal lines. As the dimensions between metal lines decrease, the edge capacitance per unit length will increase if the thickness does not scale. To reduce this capacitance and the capacitance between metal layers, low-*k* dielectric layers are being introduced. A processing example is shown in Fig. 9, and improvements will continue to be developed. The use of optical techniques to reduce delays on integrated circuits is also being studied.

The future technology challenges are still paramount, and the complexity of integrated circuits can not continue to increase at the present rate forever; however, it is very difficult to predict when Moore's law will become invalid. For example, in the early 1980s, no one could foresee the impact that chemical mechanical polishing would have on the technology. Today, new materials such as carbon fibers and other carbon structures are being investigated, which may also have a large impact in the future.

Design

VLSI chips containing over 10^9 transistors and operating at a few gigahertz (10^9 cycles/second) have been designed and fabricated, and are commercially available. Projections indicate that silicon chips containing as many as 10^{10} transistors are feasible for digital applications. The design of these new circuits starts with computer simulations that are done with computers using chips from the previous cycle. Advanced integrated circuits from current manufacturing are used to develop, design, test, and analyze the next generation of integrated circuits. *See* COMPUTER-AIDED CIRCUIT DESIGN.

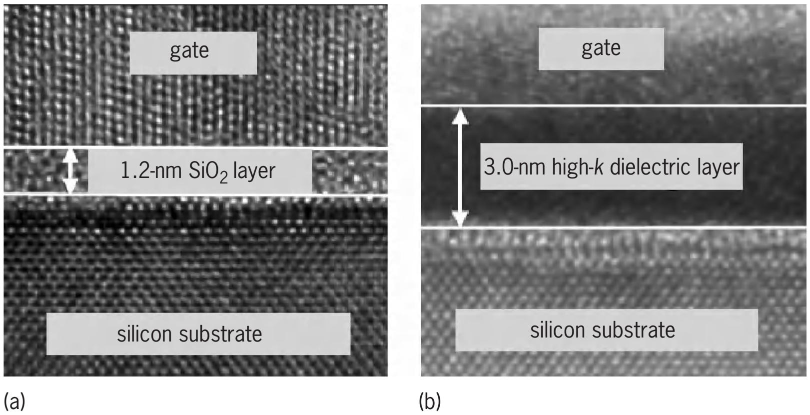


Fig. 16. **Comparison of transmission electron microscope cross sections of (*a*) standard 90-nm process chip with oxide layer and (*b*) experimental chip with thicker high-*k* (high-permittivity) dielectric. (*G. Moore, ISSCC '03, Intel Corp.*)**

The process development cycle for a new technology begins with the assumption that Moore's law continues to be valid. This assumes an approximate 30% reduction in linear dimensions and the introduction of products with the new technology 2 years after the previous generation. This does not mean that the development starts just 2 years before product introduction, for it may take more than 2 years. Some developments take many years, such as EUV lithography, which has been under development for more than 5 years (as of 2006) and may require another 5 years before it will used in manufacturing. Thus, the industry needs to anticipate technology and design requirements many years in advance to have enough time to adequately insert the new technologies and designs.

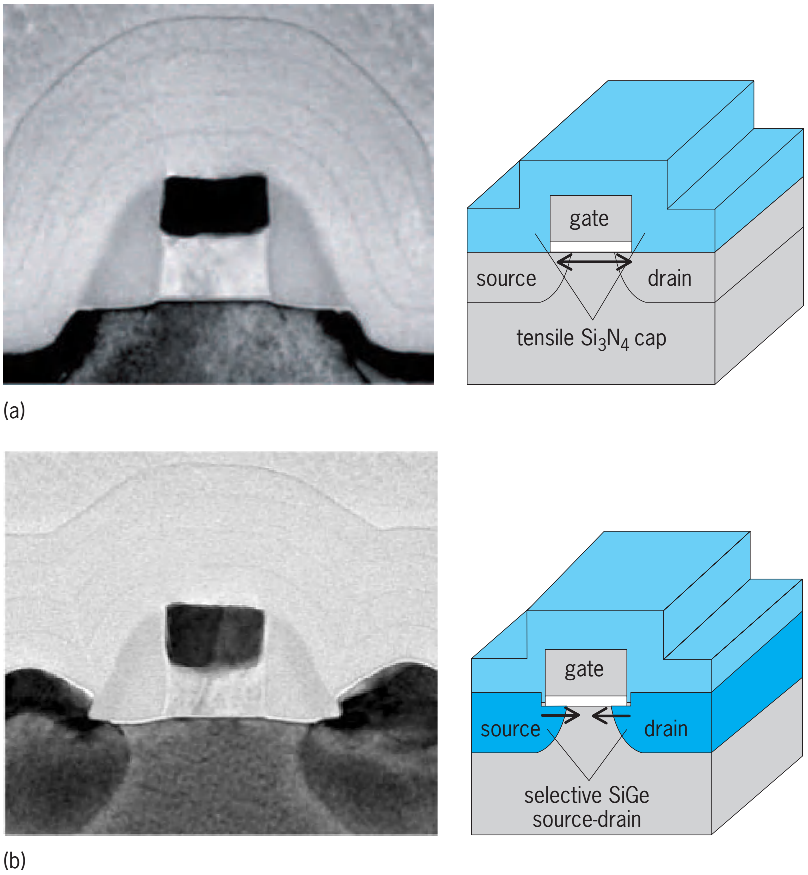


Fig. 17. **Cross-sectional views of (*a*) NMOS and (*b*) PMOS transistors with enhanced strain to improve carrier mobility. (*C.-H. Jan, Intel Corp.*)**

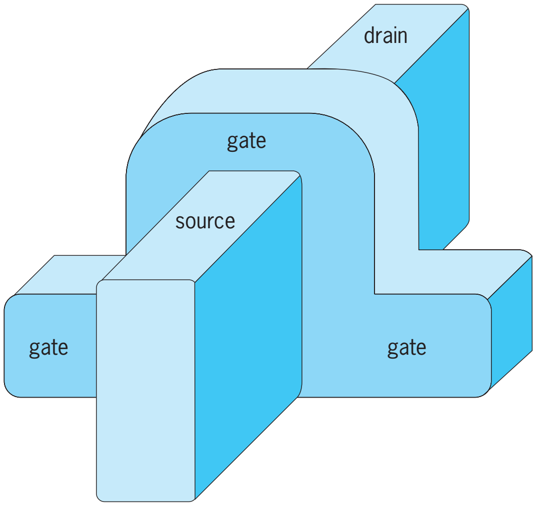


Fig. 18. Tri-gate transistor. (*R. Chau, Intel Corp.*)

The design process starts with computer simulations of the process and devices. As process flows are developed, the design rules and goals for electrical parameters become available to start the design of standard cells (smaller blocks of logic and cells), which are optimized to the design rules and performance of the new technology. A library of these cells is made, which is used by logic design teams. In technologies developed for microprocessors, SRAMs like that in Fig. 2 are often used to evaluate the process during development. There is also a lot of SRAM on microprocessors (Fig. 5), which makes it a good yield test vehicle. Chip designs from the previous technology are also used to evaluate the new technology, after shrinking the design as much as possible to match the new design features. These are often memory products (RAMs) or microprocessors, with known performance and yield indicators.

Computer-aided engineering (CAE) systems provide the environment, specific computer tools, data management, and other services that are intended to support the design of these very complex, high-performance products. In many cases, the design of complex chips requires the cooperative endeavors of large design teams; thus the CAE system must also manage the design process to ensure that proper documentation has occurred, needed changes in the design database are made, and a chosen design methodology is enforced. The design process must be adapted to the very short design cycle times from product conceptualization to manufacturing of a salable product. This is a distinctive characteristic of the semiconductor industry. *See* COMPUTER-AIDED ENGINEERING.

As it is becoming more difficult and more expensive to continue the performance improvements on the processing side, the architecture, logic, and circuit design areas become even more important to continue the overall performance improvements needed for the growth of the industry. As noted in **Fig. 19**, the computing performance as measured in MIPS has recently been going up faster than expected from just technology improvements. Some of the design features that increase overall performance include increased bit length, more memory on chip, instruction-level parallel processing (superscalar, speculative OOO), thead-level parallel processing (multithreaded), multicore parallel processing, more prediction in pipelines, and optimizing the layout to reduce delay times. In addition, special purpose

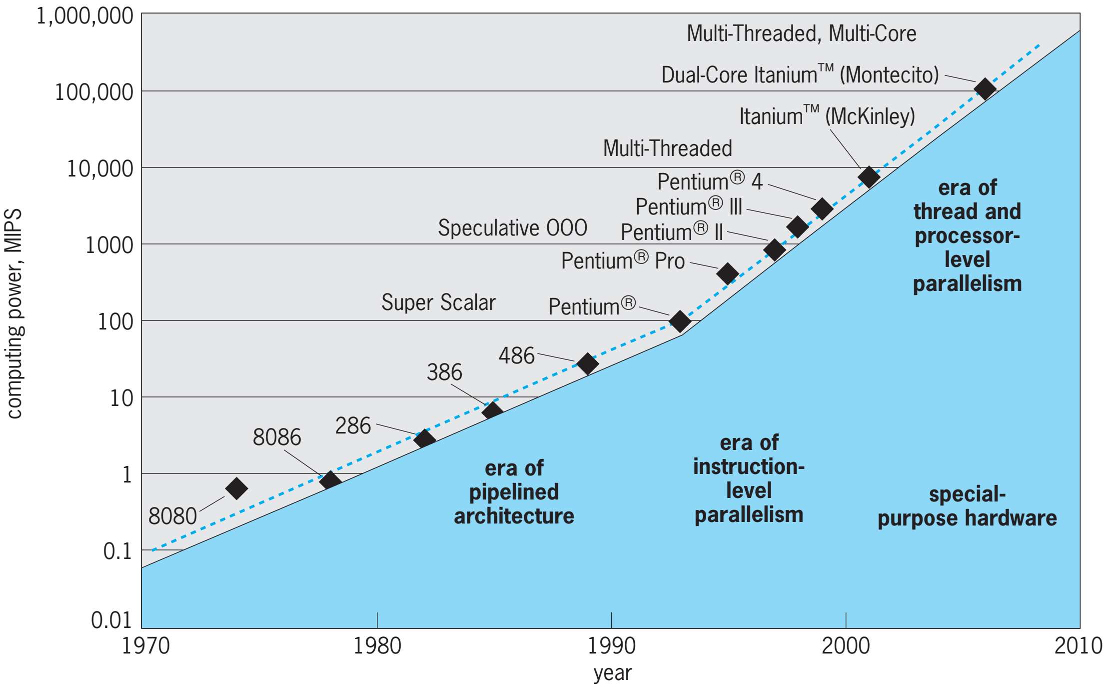


Fig. 19. Increase in computing power [measured in millions of instructions per second (MIPS)] of integrated circuits versus time and design enhancements. (*S. Borkar, Intel Corp.*)

hardware is used, which optimizes the performance of a specific function. *See* CONCURRENT PROCESSING.

Ron Burghard; Youssef El-Mansy; Neil Berglund

Bibliography. L. Berlin, *The Man Behind The Microchip*, 2005; N. G. Einspruch et al. (eds.), *VLSI Electronics: Microstructure Science*, vols. 1–24, 1981–1994; D. J. Elliott, *Integrated Circuit Fabrication Technology*, 2d ed., 1989; R. L. Geiger, P. E. Allen, and N. Strader, *VLSI Design Techniques for Analog and Digital Circuits*, 1990; P. R. Gray et al., *Analysis and Design of Analog Integrated Circuits*, 4th ed., 2001; D. A. Hodges, H. G. Jackson, and R. A. Saleh, *Analysis and Design of Digital Integrated Circuits*, 3d ed., 2004; S. I. Long and S. E. Butner, *Gallium Arsenide Digital Integrated Circuit Design*, 1990; G. Moore, *Presentation at the International Solid State Circuits Conference (ISSCC)*, 2003; J. Plummer, M. D. Deal, and P. B. Griffin, *Silicon VLSI Technology*, 2000; T. E. Price, *Introduction to VLSI Technology*, 1994; D. A. Pucknell and K. Eshraghian, *Basic VLSI Design*, 3d ed., 1994; W. S. Ruska, *Microelectronic Processing*, 1987; S. M. Sze, *Physics of Semiconductor Devices*, 2d ed., 1981; S. M. Sze, *VLSI Technology*, 2d ed., 1988; P. Van Zant, *Microchip Fabrication*, 5th ed., 2004; S. Wolf, *Silicon Processing for the VLSI Era*, vol. 4: *Deep-Submicron Process Technology*, 2002; E. S. Yang, *Microelectronic Devices*, 1988; M. Zambuto, *Semiconductor Devices*, 1989.

Integrated optics

The study of optical devices that are based on light transmission in planar waveguides, that is, dielectric structures that confine the propagating light to a region with one or two very small dimensions, on the order of the optical wavelength. The principal motivation for these studies is to combine miniaturized individual devices through waveguides or other means into a functional optical system incorporated into a small substrate. The resulting system is called an integrated optical circuit (IOC) by analogy with the semiconductor type of integrated circuit. Alternative names for an IOC include opto-electronic integrated circuit (OEIC), optical integrated circuit (OIC), and photonic integrated circuit (PIC). An integrated optical circuit could include lasers, integrated lenses, switches, interferometers, polarizers, modulators, detectors, and so forth. Important uses envisioned for integrated optical circuits include signal processing (for example, spectrum analysis and analog-to-digital conversion) and optical communications through glass or plastic fibers, which are themselves circular (or elliptical) waveguides. Integrated optical circuits could be used in such systems as optical transmitters, switches, repeaters, and receivers. *See* INTEGRATED CIRCUITS; OPTICAL COMMUNICATIONS; OPTICAL FIBERS.

The advantages of having an optical system in the form of an integrated optical circuit rather than a conventional series of components include (apart from miniaturization) reduced sensitivity to air currents and to mechanical vibrations of the separately mounted parts, low driving voltages and high efficiency, robustness, and (potentially) reproducibility and economy. As in the case of semiconductor integrated circuits, an integrated optical circuit might be fabricated on or just within the surface of one material (the substrate) modified for the different components by shaping structures (using etching, for example) or incorporating suitable substitutes or dopants, or alternatively, by depositing or expitaxially growing additional layers. This form of IOC is known as a monolithic optical integrated circuit (MOIC). It is also possible to construct independent components that are then attached to form the integrated optical circuit. This option, called hybrid, has the advantage that each component could be optimized, for example, by using gallium aluminum arsenide lasers as sources for an integrated optical circuit and silicon detectors. In the former case, the MOIC is expected to have the advantage of ease of processing, similar to the situation for monolithic semiconductor integrated circuits. The most promising materials for monolithic integrated optical circuits are direct band-gap semiconductors composed of III-V materials such as gallium aluminum arsenide (GaAlAs) and indium gallium arsenide phosphide (InGaAsP) since with suitable processing they may perform almost all necessary operations as lasers, switches, modulators, detectors, and so forth. Recently, MOICs have been fabricated by integrally bonding chips of III-V semiconductors to chips of silicon in order to combine the above optical devices with well-developed silicon electronic circuits such as CMOS (complementary metal oxide semiconductor) computer logic chips. *See* BAND THEORY OF SOLIDS; SEMICONDUCTOR.

Guided waves. The simplest optical waveguide is a three-layer or sandwich structure with the index of refraction largest in the middle or waveguiding layer (**Fig. 1**). The lower and top layers are usually the substrate and superstrate, respectively. Often, the top layer is air and the waveguide layer is referred to as a film. Sometimes, the outer regions are called cladding layers. A guided wave does not have light distributed uniformly across the waveguide, but is a pattern that depends on the indices of refraction of all three layers and the guide thickness. The waveguide is usually designed by selecting its refractive index and thickness, so that only one such characteristic pattern propagates with no change in

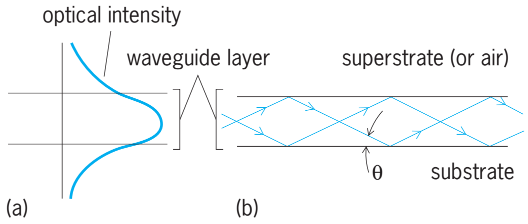


Fig. 1. Optical waveguide. (*a*) Optical intensity of the fundamental guided mode. (*b*) Ray picture of the guided mode.

shape. This pattern, referred to as the fundamental or lowest-order mode, travels down the guide with a characteristic velocity.

A mode pattern such as that of Fig. 1*a* may be viewed as the superposition of two progressive waves moving at a particular slight angle to the propagation direction as shown in Fig. 1*b*. If the ray directions are sufficiently near grazing incidence, the large refractive index of the waveguide layer will cause total internal reflection at the interfaces, and the mode will propagate along the waveguide without loss. *See* WAVEGUIDE.

Materials and fabrication. Waveguides have been made of many different materials, most of which may be categorized as ferroelectric, semiconductor, or amorphous. Examples of these classes are lithium niobate, gallium arsenide/gallium aluminum arsenide [GaAs/(GaAl)As], and glass, respectively. Methods for fabricating a waveguide layer at the surface of lithium niobate include heating the crystal in a vacuum to drive off lithium oxide, diffusing titanium metal into the crystal, or replacing lithium ions with hydrogen ions in an acid bath. These processes create a region of high refractive index near the surface; air is the superstrate. Although the refractive index of the waveguide layer decreases with depth rather than being constant, the guided mode is similar to that in Fig. 1. A semiconductor waveguide may be fabricated, for example, by successively growing crystalline layers of $(Ga_{0.7}Al_{0.3})As$, GaAs, and $(Ga_{0.7}Al_{0.3})As$. The thin GaAs waveguide layer of high refractive index is thus interposed between thicker cladding regions of the lower-index $(Ga_{0.7}Al_{0.3})As$. High-energy ion implantation of suitable dopant atoms into a semiconductor may also be used to create the high-refractive-index waveguide layer. Glass waveguides may be formed, for example, by sputter deposition of a relatively high-refractive-index glass on a lower-index glass substrate. Deposition of polymer materials also may be used to form amorphous waveguides. The advantage that polymer waveguides have over glass is that they are flexible. In recent years, waveguides and other integrated optic devices have been fabricated using polymer materials by a number of companies in order to reduce costs. Semiconductors such as GaInAsP and dielectrics such as lithium niobate are relatively expensive, and complicated processes must be used to create devices in them. The demand for inexpensive integrated optic devices has grown markedly as telecommunication networks worldwide have converted to lightwave signals transmitted over optical fibers and processed by integrated optic circuit chips. Many of the devices that formerly were made only in semiconductors or lithium niobate are now also made in polymers. These include waveguides, couplers, gratings, modulators, light detectors, light-emitting diodes (LEDs), and even lasers. In many cases, a polymer device can be substituted for one made in a semiconductor or crystalline dielectric, resulting in a significant cost reduction without excessive loss of performance. *See* FERROELECTRICS; GLASS; ION IMPLANTATION; LIGHT-EMITTING DIODE; OPTICAL MODULATORS; SEMICONDUCTOR HETEROSTRUCTURES.

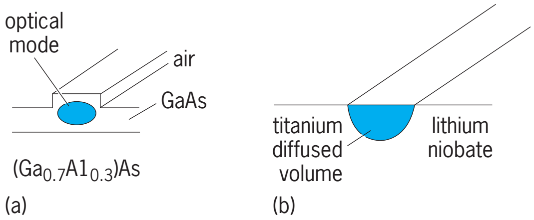


Fig. 2. **Two-dimensional transverse waveguiding. (*a*) Rib waveguide. (*b*) Diffused channel guide.**

Waveguides that confine light in two dimensions, rather than one as shown in Fig. 1, use refractive index differences in both transverse directions. Examples of these are the rib guide (**Fig. 2***a*) and the titanium-diffused channel guide (Fig. 2*b*). In fabricating a rib guide, photolithography is employed to delineate the stripe, followed by chemical or dry etching to remove the undesired material. The channel guide is produced by etching away all but a strip of metal prior to diffusion. *See* WAVEGUIDE.

Coupling of external light beams. An external light beam may be coupled into a waveguide by introducing the light at the end of the guide at an edge of the substrate or through the surface of the waveguide. The former approach may employ a lens to focus the light beam onto the guide end. Alternatively, the laser or an optical fiber is placed against, or in proximity to, the guide end, which is referred to as butt coupling. Light may also be injected through the guide surface with an auxiliary element. Simply illuminating that surface will not suffice since the refractive-index difference needed for effective waveguiding prevents light from entering at the required angle. The necessary angle, which is exactly that of the propagating waves shown in Fig. 1*b*, may be obtained by passing the light through a high-refractive-index prism whose surface is so close to the guide layer that waveguiding is disturbed in that region. The prism not only permits the introduction of light but also acts to couple light from the waveguide. For this reason, the light must illuminate the guide near the end of the prism (**Fig. 3**). The angle of light incident on the surface of the waveguide layer can also be modified to coincide with the internal angle of the propagating wave by a periodic structure or

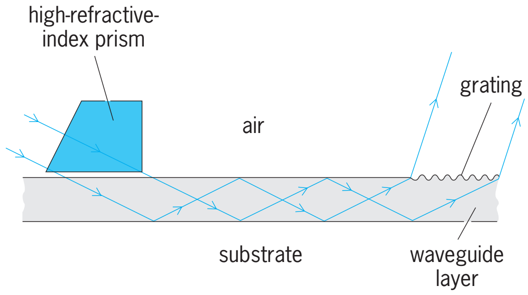


Fig. 3. **Prism input coupling and grating output coupling. With the ray directions reversed, the grating acts as an input coupler and the prism as an output coupler.**

grating on the waveguide surface. Such a grating may be made by a multiple-step process that begins by depositing a light-sensitive material called photoresist and concludes with etching and cleaning. The grating also functions as an output coupler. By reversing the incident and output beam directions in Fig. 3, the roles of the prism and grating couplers are interchanged.

Lasers and distributed feedback. The diode laser is already an integrated optics device in a sense since the lasing medium is a waveguide laser interposed between two cladding layers. The waveguide layer may be GaAs, in which case the outer regions are composed of $(Ga_{0.7}Al_{0.3})As$, for example. When used as a separate source, the crystal facets act as end reflectors. The diode laser's high-efficiency low-voltage operation, small size, and physical integrity cause it to be the laser of choice in many hybrid integrated optics applications. Moreover, it lends itself to integration in a monolithically integrated optical circuit where reflectivity may be provided by introducing an appropriate periodic structure. This could be a thickness variation in the waveguide layer. Devices utilizing such structures are called distributed feedback or distributed Bragg reflection (DBR) lasers. *See* LASER.

Switches and modulators. Both lithium niobate and gallium arsenide belong to the family of electrooptically active crystals. When an electric field is applied to these materials, their refractive indices are modified. This effect is employed in integrated optical circuit switching and modulation applications. To construct a switch, gold or other conducting electrodes are deposited on a lithium niobate integrated optical circuit surface parallel to two closely spaced waveguides. If the electrodes and waveguides are suitably designed, the applications of specific small voltages to the electrodes will cause the transfer of optical power from one waveguide to its neighbor with high efficiency and little residual power in the initial guide. Shown in **Fig. 4** is a "4 by 4" switching network in which each of four input optical signals, possibly from optical fibers butt-coupled to the waveguide inputs, may be switched to any one of the four output ports. Such an integrated optical circuit serves to interconnect four computers through optical fibers, for example. *See* ELECTROOPTICS.

A switch is in effect a modulator as well. Modulation is a process in which information is encoded onto an optical or other electromagnetic wave. Pulse modulation results simply by interrupting or connecting a lightwave in a manner intelligible to the receiver. (Morse code is an unsophisticated example.) By transferring light into or out of a waveguide in response to an electric signal at a switching electrode, the output optical wave becomes modulated; that is, the switch operates as a modulator. Devices of these types are available commercially. *See* OPTICAL MODULATORS.

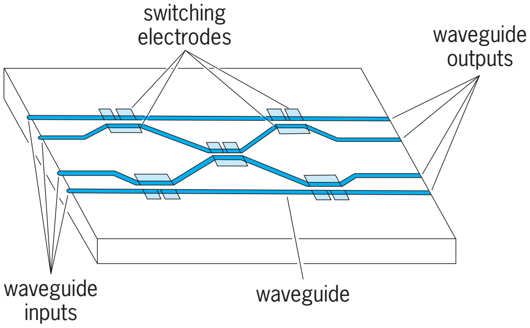


Fig. 4. A "4 by 4" directional coupler switch.

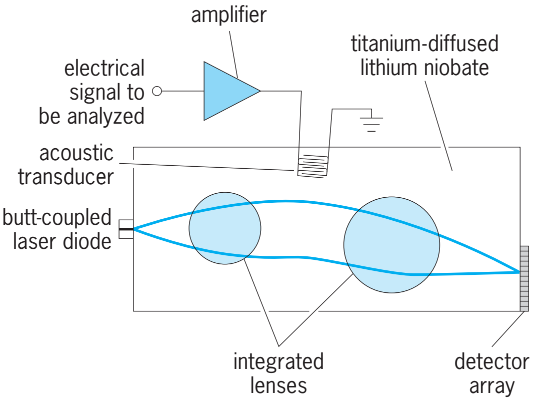


Fig. 5. Top view of an integrated optics spectrum analyzer.

Spectrum analyzer. Light propagating in an integrated optical circuit within a one-dimensional waveguide may be deflected by interacting with an acoustic wave. The deflection angle is a function of the acoustic frequency. This interaction forms the basis of an integrated optics spectrum analyzer (**Fig. 5**). An acoustic wave is generated by an electric signal whose frequencies are being analyzed. The acoustic wave interacts with a guided optical wave, and the deflected light is focused onto a detector array. The one or more detectors that respond determine the previously unknown frequencies. The integrated optical circuit incorporates lenses that focus the light and a transducer that generates the acoustic wave. The lenses may be fabricated by depositing additional high-refractive-index material in an area to decrease locally the propagation velocity of the guided wave. The transducer is made of a piezoelectric material (such as lithium niobate) which has the property of converting an electric field into a strain field. If lithium niobate is used for the integrated optical circuit, the transducer consists of a sophisticated interdigital electrode structure deposited on the waveguide surface. *See* ACOUSTOOPTICS; SPECTRUM ANALYZER.

Wavelength division multiplexing. A major application of integrated optics has taken place in the field of telecommunications, where signals are now mostly transmitted over long distances by modulated lightwaves guided by optical fibers, and OICs are often used for signal switching and processing. One of the key advantages of this technology is that a multiplicity of individual signals (telephone conversations or computer data transfers) can be transmitted simultaneously over a single optical fiber

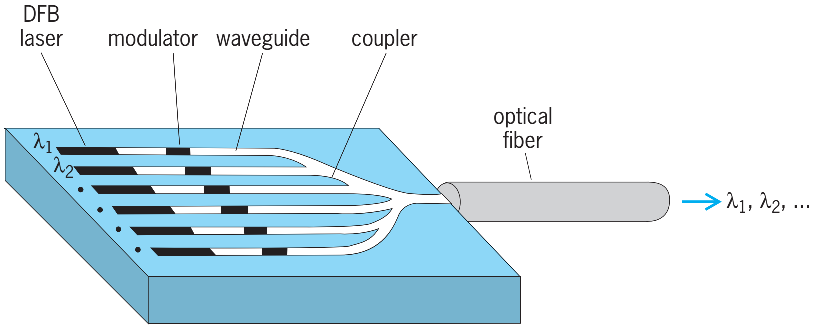


Fig. 6. A wavelength-division-multiplexed transmitter chip. DFB indicates distributed feedback.

without interfering with one another. This process is known as multiplexing. There are several different methods of multiplexing: time domain multiplexing (TDM), frequency division multiplexing (FDM), and wavelength division multiplexing (WDM). WDM is the most effective for lightwave signal transmission. The electromagnetic fields of lightwaves of different wavelength are orthogonal; that is, they do not interact. Thus, separate information channels operating at different wavelengths can be carried over a single fiber. Using increasingly sophisticated methods of WDM, the number of individual telephone channels carried by a single fiber has been increased from 1920 in 1980 to 240,000 in 2000. The signal capacity of an optical fiber is often described in terms of its bandwidth or digital data rate in gigabits per second (Gb/s). [A gigabit contains 1 billion (10^9) bits of digital information.] Currently, the better WDM systems have bandwidths of hundreds of gigabits per second transmitted over each fiber. They are called densely wavelength division multiplexed (DWDM) systems. *See* MULTIPLEXING AND MULTIPLE ACCESS.

The lightwave signals for WDM systems are usually generated by distributed feedback lasers designed to operate at different wavelengths that are accurately set by the periodicity of their reflective gratings. These lasers can be discrete devices or monolithically integrated into an OIC (**Fig. 6**). The signal information is impressed on the lightwave beam by a modulator, which can be of the electro-optic, acousto-optic, or electro-absorption type. All of these devices can be fabricated in waveguide-coupled, multilayer structures of GaAlAs or GaInAsP on a GaAs or InP substrate. At the receiving end, the multiplexed signal channels are separated by wavelength selective filters into individual waveguides, which lead to semiconductor photodiode detectors. This process is usually done by an OIC. The filters can be either gratings or array waveguides. The array waveguide (AWG) is a device that performs the spatial dispersion of light of different wavelengths in a manner similar to that of a prism. However, it is easier to fabricate in an OIC and is less sensitive to vibration than a prism. *See* FIBER-OPTIC CIRCUIT.

Photonic crystals. In both electronic integrated circuits (ICs) and optical integrated circuits (OICs), designers continually strive to decrease device size in order to increase the number of devices that can be fabricated on a single chip. High density of devices yields both improved performance and lower cost. However, electronic device density is reaching a fundamental limit because of inadequate heat dissipation. Conventional optical waveguides based on total internal reflection (TIR) are not subject to the heating effect of electrical current flowing through a wire, but they also have a fundamental limit in that tight bends in the waveguide cause excessive loss of the light. This has led researchers to investigate OICs based on photonic crystals. Photonic crystals guide light by the distributed Bragg reflection mechanism because of their periodic nature. This permits device sizes to approach that of the wavelength of the light, and 90° bends in waveguides are possible. A photonic crystal can be created in a dielectric material by drilling or etching a regularly spaced pattern of air holes that is analogous to the pattern of uniformly spaced atoms of a semiconductor crystal. Alternatively, an ordered array of columns or “fingers” can be fabricated in the dielectric by etching. The spacing of the holes or columns must be approximately the wavelength of the light in the dielectric material ($\sim$0.0001 mm). In that case, the photonic crystal responds to the guiding of light in the same way that a semiconductor crystal responds to the flow of electrons. There is an optical band gap over a range of wavelength (that is, photon energy) in which no light can propagate, just as there is a band gap in a semiconductor crystal over a range of energy in which there are no allowed electron states. It is possible to create energy levels within the photonic band gap by changing the size of some of the air holes or columns. This is the photonic equivalent of introducing dopant atoms into the semiconductor crystal. Thus, the properties of the photonic crystal can be modified to produce different devices. The critical parameters are the diameter of the air holes or columns and the contrast in refractive index. Because of the extremely small dimensions of photonic crystals, state-of-the-art semiconductor nanofabrication techniques—such as e-beam lithography, x-ray interference lithography, and ion beam etching—must be used to fabricate them. Photonic crystals have been used to make waveguides, couplers, splitters, and resonant cavities. *See* MICROLITHOGRAPHY.

Robert Hunsperger

Bibliography. J. T. Boyd (ed.), *Integrated Optics: Devices and Applications*, 1991; K. J. Ebeling, *Integrated Optoelectronics*, 1993; R. G. Hunsperger, *Integrated Optics: Theory and Technology*, 5th ed., 2002; S. I. Najafi (ed.), *Glass Integrated Optics and Optical Fiber Devices*, 1996; H. Nishihara, M. Haruna, and T. Suhara, *Optical Integrated Circuits*, 1989.

Integrated services digital network (ISDN)

A generic term referring to the integration of communications services transported over digital facilities such as wire pairs, coaxial cables, optical fibers, microwave radio, and satellites. ISDN provides end-to-end digital connectivity between any two (or more) communications devices. Information enters, passes through, and exits the network in a completely digital fashion.

Characteristics. Since the introduction of pulse-code-modulation (PCM) transmission in 1962, the worldwide communications system has been evolving toward use of the most advanced digital technology for both voice and nonvoice applications. Pulse-code modulation is a sampling technique which transforms a voice signal with a bandwidth of 4 kHz into a digital bit stream, usually of 64 kilobits per second (kbps). *See* PULSE MODULATION.

Many aspects of telecommunications are improved with digital technology. For example, digital technology lends itself to very large-scale integration (VLSI) technology and its associated benefits of miniaturization and cost reduction. In addition, computers operate digitally. Digital transport provides for human-to-human, computer-to-computer, and human-to-computer interactions. The ISDN is capable of transporting voice, data, graphics, text, and even video information over the same equipment. *See* DIGITAL COMPUTER; INTEGRATED CIRCUITS.

Service integration and standard interface. The customer has access to a wide spectrum of communications services by way of a single access link. This is in contrast to existing methods of service access, which segregate services onto specialized lines (see **illus.**).

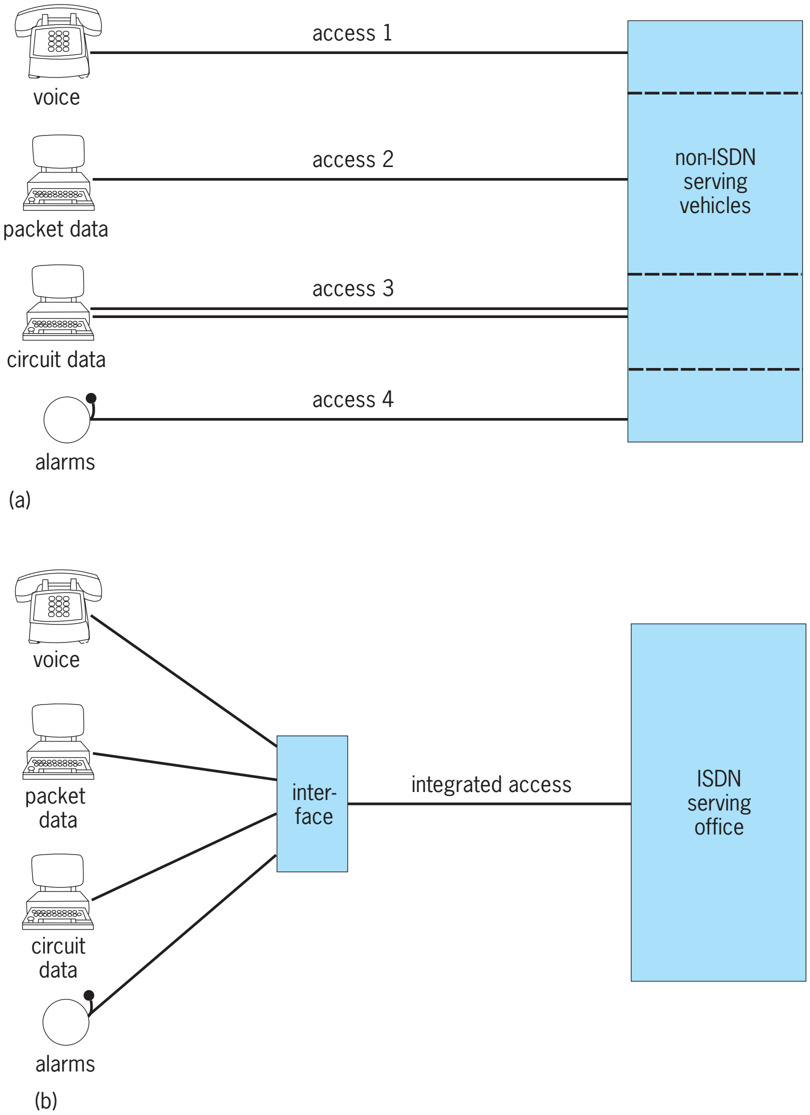
Associated with integrated access and ISDN is the concept of a standard interface. The objective of a standard interface is to allow any ISDN terminal to be plugged into any ISDN interface, resulting in terminal portability, flexibility, and ease in operation. The benefits of standard interfaces are readily visible in the United States electrical industry, where almost any electrical outlet can be used by any electrical device.

The customer's perspective of the service is through the interface. In fact, ISDN is recognized solely by the service characteristics (performance, bandwidth, and protocol) offered at the user-network boundary, rather than by any internal network components.

Signaling. ISDN provides a signaling control channel. This is a common channel, transported out of band, that provides advanced digital signaling capabilities.

Digital signaling allows computers, which do not understand voice messages, to react to the varied states which occur in a typical call. Furthermore, the time required to set up a call is reduced to the order of a few seconds, and additional messages are easily introduced when required for new service applications.

Out-of-band signaling means that the messages do not use the same channel as the customer information but are carried independently. This allows messages to be sent to the users while the call is active without interfering with the call. Thus, new services are available. For example, the identity of the calling party, sent to the called party in machine-readable form, can prevent invalid users from accessing private computers. Out-of-band signaling also enables functions to be performed at the central office much more rapidly. *See* COMPUTER SECURITY.



Comparison of segregated and integrated access. (*a*) For segregated access customer-network interfaces differ whereas (*b*) for integrated access there is one customer interface.

The signaling between switching centers in an ISDN is carried on a separate digital data network using packet-switching techniques. Not only is the call setup time decreased, but also additional information can be exchanged between originating and terminating offices for such service enhancements as caller identity and ring back.

Standards. In order to establish ISDN, it is necessary to define a set of interfaces between the customer premises equipment and the telecommunications transport network. An important goal is to make these interfaces internationally compatible. To realize this, an international team of experts in customer premises equipment and telecommunications planners has been formed under the auspices of the International Telegraph and Telephone Consultative Committee (CCITT).

Interfaces. Currently, two ISDN interfaces are defined in the CCITT, each an outgrowth of existing capabilities. Since the existing network encodes voice into a 64-kbps stream, the building block of ISDN is a transmission rate of 64 kbps.

The basic rate interface operates at 144 kbps, providing two 64-kbps channels plus a 16-kbps packet-oriented digital channel. The 64-kbps channels, referred to as B channels (for bearer channels), transport customer information, either voice or data. The 16-kbps channel, referred to as a D channel (for data channel), carries signaling and packet data.

The primary rate interface at 1.544 megabits per second (Mbps) can provide access to digital private branch exchanges (PBXs) or other large communications systems. This interface consists of twenty-four 64-kbps channels, and is an outgrowth of existing multiplexed digital carrier communication. One of these 64-kbps channels, called a D channel, operates as a signaling and packet data channel. The other twenty-three channels are B channels, as in the basic access. This interface is flexible in that several of the 64-kbps B channels can be combined to provide higher-speed services such as 384 kbps.

Additional work is under way in the CCITT to standardize bearer service subrates, such as 8, 16, and 32 kbps. Thus, a wide spectrum of customer data rates will be provided for numerous applications. ISDN is designed to provide bandwidth on demand, which means that the user will be able to specify the data rate desired on a call-by-call basis. Also, since different applications require different performance levels, performance on demand will be another key component of ISDN.

The channel designation H is used for bit rates from 384 kbps up to 45 Mbps. This is known as wideband ISDN.

As end-to-end digitalization of networks becomes commonplace at these rates, a need is expected to emerge for the ability to transmit digital information at even higher rates. Optical fiber lines have been installed that operate at 1.2×10^9 bits per second. At these rates, applications that require large amounts of information, such as high-definition real-time motion pictures or television, become practical.

Communications network. In addition to integrated-access transport from the customer to the switching node, ISDN must provide digital switching, interoffice transport, and signaling functions. Switching includes circuit switching and packet switching.

Circuit switching is used in the existing voice network where end-to-end paths are dedicated to a single call during the call duration. Packet switching is intended for data calls which, unlike voice, tend to be "bursty" in information content. Packet-switched networks bundle customer data into self-contained messages that use the transport facility only during the transmission of the packet and share the medium with packets from other users. *See* DATA COMMUNICATIONS; PACKET SWITCHING.

B channels are carried over the integrated-access line to the ISDN serving office. Here, they are routed over the circuit-switched network or are circuit-switched to a packet-switched network.

The D channel is the integrated-access control link between the customer and the ISDN serving office, ending at the office. It carries signaling information and packet data. The packet data are routed to a packet-switched network, while the signaling information is either interpreted by the local switch or sent into the network over a signaling network.

A method called asynchronous transfer mode (ATM) packages digital information to be transmitted into uniform cells of 53 bytes. These cells are then transmitted over appropriate media at will, a procedure known technically as asynchronous transmission. Each cell includes address information to guide it to its ultimate destination. Cells from diverse sources with different priorities of transmission can be mixed serially on the same media. While the basic transmitted element is a packet-like cell, switching for ATM is a mixture of packet and circuit techniques. *See* ELECTRICAL COMMUNICATIONS; SWITCHING SYSTEMS (COMMUNICATIONS).

Robert M. Wienski; Amos E. Joel, Jr.

Bibliography. G. C. Kessler, *ISDN: Concepts, Facilities, and Services*, 2d ed., 1993; B. Kumar, *Broadband Communications: A Professional's Guide to ATM, Frame Relay, SMDS, SONET, and B-ISDN*, 1995; W. Stallings, *ISDN and Broadband ISDN with Frame Relay and ATM*, 3d ed., 1994; C. K. Summers, *ISDN Implementor's Guide: Standards, Protocols and Services*, 1995.

Integration

An operation of the infinitesimal calculus which has two aspects. The roots of one go back to antiquity, for Archimedes and other Greek mathematicians used the "method of exhaustion" to compute areas and volumes. A simple example of this is the approximation to the area of a circle obtained by inscribing a regular polygon of known area, and then repeatedly doubling the number of sides. The areas of the successive polygons are computable with the help of elementary geometry. The limit of the sequence of these areas gives the area of the circle. The area of each polygon can be regarded as being made up of the sum of the areas of triangles with vertices at the center of the circle, and so the process described is a constructive definition of an integral which is the limit of a sum. Modern definitions of integrals as limits of sums are discussed in this article.

The other aspect of integration is the process of finding antiderivatives, that is, for a given function $f(x)$ to find another function $g(x)$ whose derivative is $f(x)$. The aspect is related to the first by the fundamental theorem of integral calculus, so both processes are called integration.

Isaac Newton emphasized the antiderivative aspect of integration, and his work shows how much can be done in the applications of integral calculus without introducing limits of sums. However,

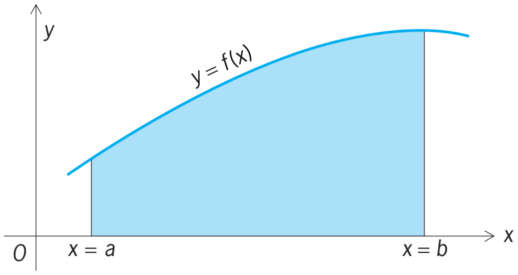


Fig. 1. Graph of *y* = *f*(*x*).

limits of sums lead to very fruitful theoretical developments in the theory of integration, as in the notion of multiple integrals, for example, and hence lead to a wider variety of applications. Leibnitz, a seventeenth-century mathematician, inspired the development in this direction, but many years elapsed before the theory was given a firm logical foundation. In the early nineteenth century A. L. Cauchy gave a clear-cut definition of the definite integral for continuous functions and a proof of its existence. Later, G. F. B. Riemann discussed the integral for discontinuous functions and gave a necessary and sufficient condition for its existence. Thus the most generally used definition of the integral as the limit of a sum has come to be called the Riemann integral.

Riemann integral. The precise definition of the Riemann integral for a real function f of one real variable x on a finite interval $a \leqq x \leqq$ may be formulated as follows. Let P be a partition of the interval $[a, b]$ into n subintervals by points t_i, where $t_{i-1} < t_i$, $t_0 = a$, $t_n = b$, and consider a sum S of the form of Eq. (1),

$$S = \sum_{i=1}^{n} f(x_i)(t_i - t_{i-1}) \tag{1}$$

where $t_{i-1} \leqq x_i \leqq t_i$. The sum S depends not only on the partition P but on the choice of the intermediate points x_i. It may happen that the sum S approaches a definite limit I when the maximum of the numbers $(t_i - t_{i-1})$ tends to zero, and in this case I is called the Riemann integral (or the definite integral) of f from a to b, and is denoted by the Leibnitzian symbol (2).

$$\int_a^b f(x)\,dx \tag{2}$$

Also, f is said to be integrable on $[a,b]$. When f is a continuous function with positive values on the interval $[a,b]$, the integral has a simple geometrical interpretation as the area bounded by the x axis, the ordinates $x = a$ and $x = b$, and the graph of $y = f(x)$ [**Fig. 1**]. It is convenient to use Eq. (3) as a definition

$$\int_b^a f(x)\,dx = -\int_a^b f(x)\,dx \tag{3}$$

of the symbol on the left.

For functions f and g which are integrable on $[a, b]$, the integral has the following properties:

(i) f is integrable on every subinterval of $[a, b]$.

(ii) For every triple of points c, d, and e in $[a, b]$,

$$\int_c^d f(x)\,dx + \int_d^e f(x)\,dx = \int_c^e f(x)\,dx$$

(iii) $f(x) + g(x)$ is integrable, and

$$\int_a^b [f(x) + g(x)]\,dx = \int_a^b f(x)\,dx + \int_a^b g(x)\,dx$$

(iv) $f(x)g(x)$ is integrable, and in particular $cf(x)$ is integrable for every real number c, and

$$\int_a^b cf(x)\,dx = c\int_a^b f(x)\,dx$$

(v) If $f(x) \leqq g(x)$ on $[a,b]$, then

$$\int_a^b f(x)\,dx \leq \int_a^b g(x)\,dx$$

(vi) $|f(x)|$ is integrable, and

$$\left|\int_a^b f(x)\,dx\right| \leq \int_a^b |f(x)|\,dx$$

It can be proved that a function f is integrable on $[a, b]$ if and only if the following two conditions are satisfied: f is bounded on $[a, b]$; and the set of points where f is discontinuous can be enclosed in a series (possibly infinite) of intervals, the sum of whose lengths is arbitrarily near to zero.

Antiderivatives. To develop the fundamental theorem of integral calculus let u be a variable in the interval $[a, b]$, on which f is integrable; then formula (4) defines a function of u which may be noted by

$$\int_a^u f(x)\,dx \tag{4}$$

$I(u)$. If f is continuous on $[a, b]$, then f is integrable, and it is also true that $I(u)$ has a derivative $I'(u) = f(u)$. Now let h be any antiderivative of f, that is, $h'(u) = f(u)$ on $[a, b]$. Then $I'(u) - h'(u) = 0$, so $I(u) - h(u) = \text{constant} = -h(a)$, by the theorem of the mean for derivatives and the fact that $I(a) = 0$, so relation (5) can be written. This is the fundamental theorem of integral calculus, and it shows that definite integrals may be calculated by the process of finding antiderivatives. For this reason antiderivatives are frequently called indefinite integrals and denoted by $\int f(x)\,dx$, and special methods of finding indefinite integrals for frequently occurring functions occupy a large part of elementary calculus. The principal methods are outlined in the next section. The standard notation for an indefinite integral of $f(x)$ is formula (6).

$$I = I(b) = h(b) - h(a) \tag{5}$$

$$\int f(x)\,dx \tag{6}$$

Elementary methods of integration. Obviously, each formula for differentiation yields a formula for indefinite integration.

Method of substitution. From these indefinite integrals, additional formulas can be obtained by the method of substitution, which is based on the chain rule for differentiating composite functions. For example, from formula (7) one obtains Eq. (8), by the substitution $v = a^2 + x^2$, and Eq. (9), by the substitution $v = \sin x$.

$$\int x^m dx = \frac{x^{m+1}x}{m+1}, m \neq -1 \tag{7}$$

$$\int (a^2 + x^2)^m x\, dx = \frac{(a^2+x^2)^{m+1}}{2(m+1)}, m \neq -1 \tag{8}$$

$$\int \sin^m x \cos x\, dx = \frac{\sin^{m+1} x}{m+1}, m \neq -1 \tag{9}$$

Method of partial fractions. In case $f(x)$ is a quotient of two polynomials in x, $f(x)$ may be represented as a sum of a polynomial and terms of the form, shown as notations (10) and (11), where m is a positive in-

$$\frac{1}{(ax+b)^m} \qquad a \neq 0 \tag{10}$$

$$\frac{Ax+B}{(cx^2+dx+e)^m} \qquad c \neq 0, d^2 - 4ce < 0 \tag{11}$$

teger and the coefficients are real. This is called the method of partial fractions. Its actual execution requires a knowledge of the factors of the denominator of $f(x)$. Terms of the form (10) have known indefinite integrals. Terms of the form (11) with $m = 1$ can also be integrated by elementary formulas. When $m > 1$, a term of the form (11) can be written as a sum of terms of the forms (12) and (13),where $\alpha =$

$$\frac{\alpha}{(\gamma x^2 + dx + e)^{m-1}} \tag{12}$$

$$\frac{(\beta x + \gamma)(2cx + d)}{(cx^2 + dx + e)^m} \tag{13}$$

$(2dA - 4cB)/(d^2 - 4ce)$, $\beta = -\alpha/2$ and $\gamma = (dB - 2eA)/(d^2 - 4ce)$. By the process of integration by parts, the indefinite integral of a term of the form (13) can be expressed in terms of the integral of a term such as (11), with m replaced by $m - 1$. Reduction formulas of this type are given in standard tables of integrals.

A rational function of $\sin x$ and $\cos x$ can be integrated by the method just described after the substitution $u = \tan (x/2)$ has been applied. This substitution reduces the problem to the integration of a rational function of u. Various functions involving radicals can be integrated by means of trigonometric substitutions or by other substitutions which reduce the problem to the integration of a rational function. For example, by setting $u = a + bx$, one finds Eq. (14),

$$\int x^m (a+bx)^p\, dx = \int \left(\frac{u-a}{b}\right)^m u^p \frac{du}{b} \tag{14}$$

and when m is a nonnegative integer the right side can be multiplied out and integrated by the power formula, even though p is a fraction.

Integration by parts. This is a very powerful and important method. It follows from the formula for differentiating a product, namely Eq. (15). After the terms

$$d(fg) = f\, dg + g\, df \tag{15}$$

are rearranged and integrated, Eq. (16) or (17) may

$$\int f\, dg = fg - \int g\, df \tag{16}$$

$$\int fg'\, dx = f(x)g(x) - \int gf'\, dx \tag{17}$$

be written, where the arbitrary constant, as usual, has been omitted. As an example, consider Eq. (18),

$$I = \int \frac{2x^2 dx}{(x^2+1)^m} \tag{18}$$

which is a special case of the form (13). Let $f = x$, $g' = 2x/(x^2 + 1)^m$. Then $f' = 1$, and g can be written as Eq. (19), and by Eq. (17), Eq. (20) follows.

$$g = \frac{1}{(1-m)(x^2+1)^{m-1}} \tag{19}$$

$$I = \frac{x}{(1-m)(x^2+1)^{m-1}} - \frac{1}{(1-m)} \int \frac{dx}{(x^2+1)^{m-1}} \tag{20}$$

The indefinite integrals of many of the commonly occuring functions are given in tables of integrals in handbooks and textbooks.

The elementary functions are those expressible by means of a finite number of algebraic operations and trigonometric and exponential function and their inverses. The integrals of many elementary functions are known to be not elementary, so they define new functions. A number of these are sufficiently important that their values have been tabulated. Examples are Eqs. (21) and (22). $F(k, x)$ is called an elliptic inte-

$$F(k,x) = \int_0^x \frac{dx}{\sqrt{(1-x^2)(1-k^2x^2)}} \qquad (k^2 \neq 1) \tag{21}$$

$$\mathrm{Si}\, x = \int_0^x \frac{\sin x\, dx}{x} \tag{22}$$

gral of the first kind. The inverse of $u = F(k, x)$ is an elliptic function called the sine amplitude of u and denoted by $x = \mathrm{sn}\, u$. Si x is called the sine integral. Some nonelementary functions have been included in tables of integrals. *See* ELLIPTIC FUNCTION AND INTEGRAL.

Improper integrals. This term is used to refer to an extension of the notion of definite integral to cases in which the integrand is unbounded or the domain of integration is unbounded. Consider first the case when $f(x)$ is integrable (in the sense defined above) and hence bounded on every interval $[a + \epsilon, b]$ for $\epsilon > 0$, but is unbounded on $[a,b]$. Then by definition, Eq. (23) is written, provided the limit on the right

$$\int_a^b f(x)\, dx = \lim_{\epsilon \to 0} \int_{a+\epsilon}^b f(x)\, dx \tag{23}$$

exists. For example, if $f(x) = x^{-2/3}$, relation (24) follows. Similarly, Eq. (25) follows.

$$\int_0^1 x^{-2/3}\,dx = \lim_{\epsilon\to 0}\int_\epsilon^1 x^{-2/3}\,dx = \lim_{\epsilon\to 0} 3(1-\epsilon^{1/3}) = 3 \qquad (24)$$

$$\int_{-1}^0 x^{-2/3}\,dx = \lim_{\epsilon\to 0}\int_{-1}^{-\epsilon} x^{-2/3}\,dx = \lim_{\epsilon\to 0} 3[(-\epsilon)^{1/3}+1] = 3 \qquad (25)$$

When $f(x)$ is integrable on every finite subinterval of the real axis, by definition Eqs. (26) and (27) hold, provided the limits on the right exist.

$$\int_a^{+\infty} f(x)\,dx = \lim_{b\to+\infty}\int_a^b f(x)\,dx \qquad (26)$$

$$\int_{-\infty}^a f(x)\,dx = \lim_{b\to-\infty}\int_b^a f(x)\,dx \qquad (27)$$

More general cases are treated by dividing the real axis into pieces, each of which satisfies one of the conditions just specified. Equation (28) is an example.

$$\int_{-\infty}^{+\infty} x^{-5/3}\,dx = \lim_{a\to-\infty}\int_a^{-1} x^{-5/3}\,dx + \lim_{\delta\to 0}\int_{-1}^{-\delta} x^{-5/3}\,dx + \lim_{\epsilon\to 0}\int_\epsilon^1 x^{-5/3}\,dx + \lim_{b\to+\infty}\int_1^b x^{-5/3}\,dx \qquad (28)$$

Because the second and third of the limits on the right do not exist, integral (29) does not exist.

$$\int_{-\infty}^{+\infty} x^{-5/3} dx \qquad (29)$$

However, it is sometimes useful to assign a value to it, called the Cauchy principal value, by replacing definition (28) by Eq. (30).

$$\int_{-\infty}^{+\infty} x^{-5/3}\,dx = \lim_{e\to 0}\left(\int_{-1}^{-\epsilon} x^{-5/3}\,dx + \int_\epsilon^1 x^{-5/3}\,dx\right) + \lim_{b\to+\infty}\left(\int_{-b}^{-1} x^{-5/3}\,dx + \int_1^b x^{-5/3}\,dx\right) \qquad (30)$$

In this particular case the value is zero.

The preceding definitions apply to cases when the integrand $f(x)$ is bounded, except in arbitrarily small neighborhoods of a finite set of points. In the closing years of the nineteenth century various extensions were made to more general cases. Then Henri Lebesgue produced a comparatively simple general theory for the case of absolutely convergent integrals. The integral of Lebesgue will be discussed below. The integral of Denjoy includes both nonabsolutely convergent integrals and the integral of Lebesgue.

Multiple integrals. The concept called the Riemann integral can be extended to functions of several variables. The case of a function of two variables illustrates sufficiently the additional features which arise. To begin with, let $f(x, y)$ denote a real function defined on a rectangle of the form (31).

$$R\colon a \le x \le b \qquad c \le y \le d \qquad (31)$$

Let P be a partition of R into n nonoverlapping rectangles R_i with areas A_i, and let (x_i, y_i) be a point of R_i. Define S by Eq. (32).

$$S = \sum_{i=1}^{n} f(x_i, y_i)A_i \qquad (32)$$

In case the sum S tends to a definite limit I when the maximum diagonal of a rectangle R_i tends to zero, then f is said to be integrable over R, and the limit I is called the Riemann integral of f over R. It will be denoted here by the abbreviated symbol $\int\int_R f$. Properties corresponding to those numbered (i) to (iv) can be proved for these double integrals, except that property (ii) should now read as follows:

(ii′) If the rectangle R is the union of two rectangles S and T, then Eq. (33) can be written.

$$\int\int_R f = \int\int_S f + \int\int_T f \qquad (33)$$

Necessary and sufficient conditions for a function $f(x, y)$ to be integrable over R can be stated as follows: f is bounded on R; and the set of points where f is discontinuous can be enclosed in a series of rectangles, the sum of whose areas is arbitrarily near to zero.

To define the integral of a function $f(x, y)$ over a more general domain D where it is defined, suppose D is enclosed in a rectangle R, and define $F(x, y)$ by Eqs. (34). Then f is integrable over D if case F is integrable over R, and Eq. (35) holds, by definition.

$$\begin{aligned} F(x,y) &= f(x,y) \text{ in } D \\ F(x,y) &= 0 \text{ outside } D \end{aligned} \qquad (34)$$

$$\int\int_D f = \int\int_R F \qquad (35)$$

When the function $f(x, y)$ is continuous on D, and D is defined by inequalities of the form (36), where the functions $\alpha(x)$ and $\beta(x)$ are continuous, the double integral of f over D always exists, and may be represented in terms of two simple integrals by Eq. (37).

$$\begin{gathered} a \le x \le b \\ \alpha(x) \le y \le \beta(x) \end{gathered} \qquad (36)$$

$$\int\int_D f = \int_a^b \left[\int_{\alpha(x)}^{\beta(x)} f(x, y)\,dy\right] dx \qquad (37)$$

In many cases, this formula makes possible the evaluation of the double integral.

Improper multiple integrals have been defined in a variety of ways. There is not space to discuss these definitions and their relations here.

Line, surface, and volume integrals. A general discussion of curves and surfaces requires an extended treatise. The following outline is restricted to the simplest cases.

A curve may be defined as a continuous image of an open interval. Thus a curve C in the xy plane is given by a pair of continuous functions of one variable, Eq. (38), whereas a curve in space is given by

$$x = f(u) \quad y = g(u) \quad a < u < b \tag{38}$$

a triple of such functions. The mapping (38) is more properly called a representation of a curve, and a curve is then defined as a suitable class of such representations. Note that a curve such as Eq. (38) is a path rather than a locus. For example, notation (39) is the path going twice around the unit circle, Eq. (40),

$$x = \cos u \quad y = \sin u \quad 0 < u < 4\pi \tag{39}$$

$$x^2 + y^2 = 1 \tag{40}$$

in the counterclockwise direction. A curve is called smooth in case the function f and g have continuous derivatives $f'(u)$ and $g'(u)$ which are never simultaneously zero.

The length of a smooth curve C may be defined by Eq. (41) when this integral exists (as a proper or

$$L(C) = \int_a^b \sqrt{[f'(u)]^2 + [g'(u)]^2 du} \tag{41}$$

improper integral). When the integral does not exist, the length of C is said to be infinite. This definition of length in terms of the integral may be shown to coincide with the geometric definition in a much larger class than the class of smooth curves. When C has finite length, the position vector $[f(u), g(u)]$ approaches definite limits when u tends to a or to b, which are the ends of the curve C. A smooth curve has at every point a nonzero tangent vector $\mathbf{T}(u)$, with components $f'(u)$ and $g'(u)$, so that the integrand of $L(C)$ is the length $|\mathbf{T}(u)|$ of this vector. Thus one may write $ds = |\mathbf{T}(u)|\ du$, where s is the arc length measured from some convenient point on C.

Let $A(x, y)$ be a bounded continuous function defined on a set containing a plane curve C having finite length. Then A determines three functions of C called line integrals, whose symbols and definitions, Eqs. (42), follow. If $-C$ denotes the curve

$$\begin{aligned} \int_C A\ dx &= \int_a^b A[f(u), g(u)] f'(u)\ du \\ \int_C A\ dy &= \int_a^b A[f(u), g(u)] g'(u)\ du \\ \int_C A\ ds &= \int_a^b A[f(u), g(u)]\ |T(u)|\ du \end{aligned} \tag{42}$$

C traversed in the opposite direction, as shown in Eqs. (43), a reversal of the orientation of C changes

$$\begin{aligned} \int_{-C} A\ dx &= -\int_C A\ dx \\ \int_{-C} A\ dy &= -\int_C A\ dy \\ \int_C A\ ds &= \int_C A\ ds \end{aligned} \tag{43}$$

the signs of the first two but not of the third of these functions.

The extension of the preceding definitions to curves in three-space is made in an obvious way. An important application of line integrals is to express the work done by a force field on a moving particle. Thus, if a force field $\mathbf{F}(x, y, z)$ has components $F_1(x, y, z)$, $F_2(x, y, z)$, and $F_3(x, y, z)$ in the directions of the x, y, and z axes, then the work done by the field on a particle moving on a curve C is given by formula (44). When $\mathbf{T}$ is the tangent vector of C, expressed

$$W = \int_C F_1 dx + F_2 dy + F_3 dz \tag{44}$$

in terms of a parameter u, the work W can be expressed in terms of the dot product by Eq. (45). If $\mathbf{T}_1$ denotes the tangent vector of unit length, notation (46) then can be written.

$$W = \int_a^b \mathbf{F} \cdot \mathbf{T}\ du \tag{45}$$

$$W = \int_C \mathbf{F} \cdot \mathbf{T}_1\ ds \tag{46}$$

To pass from curves to surfaces replace the open interval $a < u < b$ by a bounded connected open set D in a uv plane, which may be called the parameter plane. The domain D can be restricted to be of sufficiently simple shape so that every function which is continuous and bounded in D is integrable over D, as a multiple integral. For example, D may be the interior of a circle, or of a rectangle. Then a surface S in three-space is defined by a triple of functions continuous on D, Eqs. (47). Such a surface is called smooth

$$x = f(u, v) \quad y = g(u, v) \quad z = h(u, v) \tag{47}$$

in case the functions f, g, and h have continuous first partial derivatives in D, and the three Jacobians listed in formulas (48), are never simultaneously zero. A

$$J_1 = \frac{\partial(y, z)}{\partial(u, v)} \quad J_2 = \frac{\partial(z, x)}{\partial(u, v)} \quad J_3 = \frac{\partial(x, y)}{\partial(u, v)} \tag{48}$$

smooth surface has at every point a nonzero normal vector $\mathbf{J}(u, v)$ with components J_1, J_2, J_3, and length $|\mathbf{J}|$.

The area of a smooth surface S may be defined by the integral (49), and it is always finite when the

$$\sigma(S) = \int\int_D |\mathbf{J}(u, v)| \tag{49}$$

vector $\mathbf{J}$ has bounded length.

If $A\ (x, y, z)$ is a bounded continuous function defined on a set containing a surface S with finite area, four surface integrals can be defined as in notation

(50). If $\mathbf{F}(x,y,z)$ is a vector field with components

$$\begin{aligned}
&\int\!\!\int_S A\, dy\, dz = \\
&\qquad \int\!\!\int_D A[f(u,v), g(u,v), h(u,v)]J_1(u,v) \\
&\int\!\!\int_S A\, dz\, dx = \\
&\qquad \int\!\!\int_D A[f(u,v), g(u,v), h(u,v)]J_2(u,v) \\
&\int\!\!\int_S A\, dx\, dy = \\
&\qquad \int\!\!\int_D A[f(u,v), g(u,v), h(u,v)]J_3(u,v) \\
&\int\!\!\int_S A\, d\sigma = \\
&\qquad \int\!\!\int_D A[f(u,v), g(u,v), h(u,v)]|\mathbf{J}(u,v)
\end{aligned} \tag{50}$$

F_1, F_2, F_3, and $\mathbf{n}$ denotes the unit vector normal to S, with components $J_1/|\mathbf{J}|, J_2/|\mathbf{J}|, J_3/|\mathbf{J}|$, then a combination, formula (51), of the first three kinds of surface

$$\int\!\!\int_S F_1\, dy\, dz + F_2\, dz\, dx + F_3\, dx\, dy \tag{51}$$

integral may also be written in the form of expression (52),where $\mathbf{F} \cdot \mathbf{n}$ is the dot product. Expression

$$\int\!\!\int_S \mathbf{F} \cdot \mathbf{n}\, d\sigma \tag{52}$$

(51) or (52) is referred to as the integral of the vector $\mathbf{F}$ over the side of the surface S to which the vector $\mathbf{n}$ points. The integral over the opposite side of S is denoted by formula (53), whose value is the negative of that in (51), by Eq. (54). The last integral in

$$\int\!\!\int_S F_1\, dz\, dy + F_2\, dx\, dz + F_3\, dy\, dx \tag{53}$$

$$\frac{\partial(z,y)}{\partial(u,v)} = -J_1, \text{ and so on} \tag{54}$$

Eq. (50), however, is independent of a choice of side of the surface.

The important formulas of Stokes and of Gauss, Eqs. (55), are expressed in terms of line and surface

$$\begin{aligned}
\int_C \mathbf{F} \cdot \mathbf{T}\, ds &= \int\!\!\int_S \text{curl}\, \mathbf{F} \cdot \mathbf{n}\, d\sigma \quad \text{(Stokes)} \\
\int\!\!\int_S \mathbf{F} \cdot \mathbf{n}\, d\sigma &= \int\!\!\int\!\!\int_V \text{div}\, \mathbf{F} \quad \text{(Gauss)}
\end{aligned} \tag{55}$$

integrals and the triple or volume integral analogous to the double integral. In the formula of Stokes, $\mathbf{F}$ is a vector field, S is a smooth surface having a piecewise smooth boundary curve C, directed so that S lies to the left of an observer proceeding along C on the side of S on which the unit normal vector $\mathbf{n}$ is chosen, and $\mathbf{T}$ is the unit tangent vector to C. In terms of the components of F, Stokes' formula may be written as Eq. (56), where $F_{3y} = \partial F_3/\partial y$, and so on. When the

$$\begin{aligned}
\int_C F_1\, dx + F_2\, dy + F_3\, dz &= \int\!\!\int_S (F_{3y} - F_{2z})\, dy\, dz \\
&+ (F_{1z} - F_{3x})\, dz\, dx + (F_{2x} - F_{1y})\, dx\, dy
\end{aligned} \tag{56}$$

surface S lies in the xy plane, this equation reduces to Green's formula, Eq. (57). The latter is readily proved

$$\int_C F_1 dx + F_2 dy = \int\!\!\int_S (F_{2x} - F_{1y})\, dx\, dy \tag{57}$$

for regions S of simple shape, and from it Stokes' formula may be deduced.

In the formula of Gauss (also called the divergence theorem), V is a space domain bounded by a smooth or piecewise smooth surface S, and $\mathbf{n}$ is the exterior normal of S. In terms of the components of F, Gauss' formula may be written as Eq. (58).

$$\begin{aligned}
&\int\!\!\int_S F_1\, dy\, dz + F_2\, dz\, dx + F_3\, dx\, dy \\
&\qquad = \int\!\!\int\!\!\int_V (F_{1x} + F_{2y} + F_{3z})
\end{aligned} \tag{58}$$

See CALCULUS OF VECTORS.

Functions defined by integrals. Mention has already been made of elliptic functions, which may be defined as the inverse functions of elliptic integrals. Certain definite integrals may also be used to define important nonelementary functions which occur frequently in applications. For example, the gamma function may be defined by Eq. (59) for $x > 0$. The

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1}\, dt \tag{59}$$

Bessel functions of integral order may be defined by Eq. (60) for all values of x. The properties of these

$$J_n(x) = \frac{1}{\pi} \int_0^\pi \cos\, (nt - x\, \sin t)\, dt \tag{60}$$

functions may be derived from these expressions with the help of various methods of advanced analysis, such as the theory of functions of a complex variable. *See* BESSEL FUNCTIONS; GAMMA FUNCTION.

In general, if $f(x, t)$ is an integrable function of t on $a \leqq t \leqq b$ for each x, then Eq. (61) is a well-defined function of x. When suitable conditions are

$$g(x) = \int_a^b f(x,t)\, dt \tag{61}$$

satisfied by the function f, the derivative $g'(x)$ may be calculated by formula (62). This process is called

$$g'(x) = \int_a^b \frac{\partial}{\partial x} f(x,t)\, dt \tag{62}$$

differentiation under the integral sign. It is valid in case f and $\partial f/\partial x$ are continuous in (x,t) for $c \leqq x \leqq d$, $a \leqq t \leqq b$, where a and b are finite, and even in more

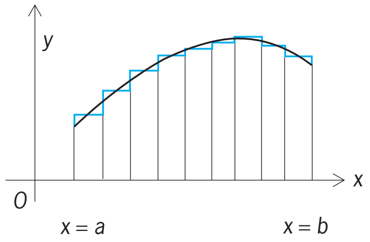


Fig. 2. Division of interval [*a*, *b*] into parts.

general cases. However, in the simple example given in Eqs. (63), formula (62) leads to a wrong result.

$$\begin{aligned} f(x,t) &= x^3 e^{-tx^2} \\ g(x) &= \int_0^\infty f(x,t)\,dt \end{aligned} \tag{63}$$

Approximate and mechanical integration. The definition of the definite integral itself gives a means of calculating its value approximately. There remains the question of a suitable selection of the functional values $f(x_i)$ for use in formula (1). It is usual to divide the interval $[a, b]$ into a number n of equal parts of length h (**Fig. 2**). If the points x_i are taken at the midpoints of the subintervals, one obtains the midpoint formula, Eq. (64). For a curve that is concave downward, this gives too large a value.

$$S = h\sum_{i=1}^{n} f[a+(i-1/2)h] \tag{64}$$

Another formula, called the trapezoidal rule, is derived by calculating the area below a polygon inscribed in the graph of $f(x)$ [**Fig. 3**]. This formula is Eq. (65). It gives too small a value for a curve that is concave downward.

$$T = \frac{h}{2}\left[f(a)+f(b)+2\sum_{i=1}^{n-1} f(a+ih)\right] \tag{65}$$

The error in the approximation may sometimes be reduced without increasing the number of subintervals by use of the parabolic rule (Simpson's rule). To obtain the formula, the subarcs of the graph of $f(x)$ are replaced by arcs of parabolas rather than by line segments. Since three points determine a parabola (with vertical axis), the interval $[a, b]$ is divided into an even number n of subintervals. The area under the parabola passing through the points on the graph of $f(x)$ having abscissas given in expressions (66) is given by expression (67). Therefore, the parabolic rule gives approximation (68).

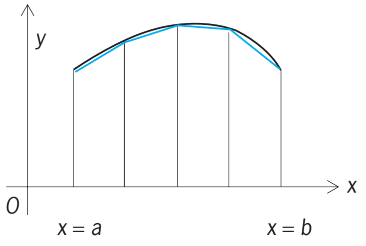


Fig. 3. Polygon inscribed in graph of *f*(*x*).

$$a+(2i-2)h \quad a+(2i-1)h \quad a+2ih \tag{66}$$

$$\frac{h}{3}\{f[a+(2i-2)h] + 4f[a+(2i-1)h] + f(a+2ih)\} \tag{67}$$

$$P = \frac{h}{3}\left\{f(a)+f(b)+2\sum_{i=1}^{(n/2)-1} f(a+2ih) + 4\sum_{i=1}^{n/2} f[a+(2i-1)h]\right\} \tag{68}$$

When a function $f(x)$ is given only by a table, its integral is most simply computed by one of these formulas or a similar formula.

When a function $f(x)$ is given by a graph, mechanical means may be used to calculate associated areas. One such means is the polar planimeter, which registers on a rotating wheel the area enclosed by a closed curve around which a tracing point is passed. The integraph, invented by Abdank Abakanowicz, is designed to draw the graph of an indefinite integral of $f(x)$ when a tracing point is passed over the graph of $f(x)$. These simple devices were the forerunners of more complex machines designed to solve differential equations, such as the differential analyzer of Vannevar Bush. For large-scale computations involving formulas such as (64), (65), or (68), a digital computer may be preferred.

Other methods of integration. The method of differentiation under the integral sign is often convenient for the evaluation of definite integrals, even when other methods are available. For example, Eq. (69) leads to Eq. (70). Or, if one has Eq. (71), Eq. (72) follows.

$$g(x) = \int_0^a e^{xt}\,dt = \frac{e^{xa}-1}{x} \tag{69}$$

$$g'(x) = \int_0^a te^{xt}\,dt = \frac{ae^{xa}}{x} - \frac{e^{xa}-1}{x^2} \tag{70}$$

$$g(x) = \int_0^a \frac{dt}{t^2+x^2} = \frac{1}{x}\arctan(a/x) \tag{71}$$

$$\begin{aligned} g'(x) &= -2x\int_0^a \frac{dt}{(t^2+x^2)^2} \\ &= -\frac{1}{x^2}\arctan(a/x) - \frac{a}{x(a^2+x^2)} \end{aligned} \tag{72}$$

Another example in which the conditions for validity of the method still hold, though they are more difficult to verify, is given by Eq. (73), by means of which Eq. (74) can be derived,

$$g(x) = \int_0^\infty e^{-xt}\frac{\sin t}{t}\,dt \qquad 0 \le x < \infty \tag{73}$$

$$g'(x) = -\int_0^\infty e^{-xt}\sin t\,dt = -\frac{1}{1+x^2} \tag{74}$$

by integration by parts twice. Hence $g(x) = -\arctan x + C$, and $C = \pi/2$, since Eq. (75) can be proved. In particular, Eq. (76) holds.

$$\lim_{x\to\infty} g(x) = 0 \qquad (75)$$

$$g(0) = \int_0^\infty \frac{\sin t}{t} dt = \frac{\pi}{2} \qquad (76)$$

In case an indefinite integral is not elementary but the integrand is representable by an infinite series, the method of integrating term by term gives a representation for the integral which may be used for purposes of approximation. This can be seen in Eq. (77).

$$\mathrm{Si}\,(x) = \int_0^x \frac{\sin t\; dt}{t} \sum_0^\infty \frac{(-1)^n x^{2n+1}}{(2n+1)^2(2n)!} \qquad (77)$$

Certain definite integrals may be readily evaluated by means of contour integrals, that is, integrals of analytic functions taken along curves in the complex plane.

Integral of Lebesgue. In the study of the "space" of real functions defined, for example, on the interval $a \leqq x \leqq b$, it is frequently useful to take formula (78) as the distance between the functions f and g.

$$\int_a^b |f(x) - g(x)|\, dx \qquad (78)$$

This distance already has a meaning when f and g are Riemann-integrable, that is, bounded and not too discontinuous, in the sense specified for the Riemann integral. There is no generally useful extension of the concept of integral to apply to all real functions on $[a, b]$, but it is desirable to extend it to apply to the functions obtained from the continuous ones by certain limiting processes. In particular, it is desirable to have correspond to each sequence (f_n) of functions satisfying the Cauchy condition for convergence in terms of the distance (78), namely Eq. (79),

$$\lim_{\substack{m\to\infty\\ n\to\infty}} \int_a^b |f_m(x) - f_n(x)|\, dx = 0 \qquad (79)$$

a function g which is integrable (in the extended sense), and for which Eq. (80) holds. An extended

$$\lim_{n\to\infty} \int_a^b |f_n(x) - g(x)|\, dx = 0 \qquad (80)$$

definition of integral having this property was given by H. L. Lebesgue in his thesis. It made obsolete many of the complicated extensions of the Riemann integral which had been previously proposed. Following Lebesgue, various mathematicians have proposed other ways of defining the integral which are equivalent to that of Lebesgue.

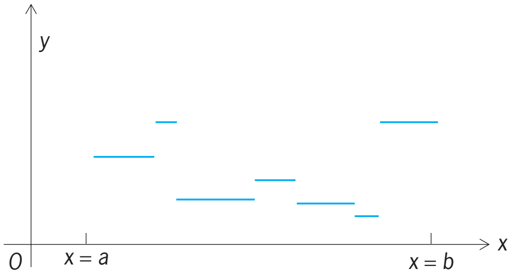


Fig. 4. Step function *f*(*x*).

F. Riesz devised a definition which can be stated quite simply, at least for the case of a bounded function $g(x)$. As a first definition, a point set S in the interval $[a, b]$ has measure zero in case it can be enclosed in a sequence (finite or infinite) of intervals, the sum of whose lengths is arbitrarily small. Also, a step function $f(x)$ is defined as one which is constant on each interval of a partition of $[a, b]$, as in **Fig. 4**. The Riemann integral of a step function is expressible as a finite sum. Then a bounded function $g(x)$ is integrable in Lebesgue's sense in case it is the limit of a uniformly bounded sequence of step functions $f_n(x)$, at each point of $[a, b]$ except those in a set S with measure zero, and by definition Eq. (81) can be

$$\int_a^b g(x)\, dx = \lim_{n\to\infty} \int_a^b f_n(x)\, dx \qquad (81)$$

written. In case the step functions f_n in Eq. (81) are replaced by Lebesgue-integrable functions forming a uniformly bounded sequence, no new functions g are obtained.

The integral of Lebesgue is also defined for unbounded functions, but the points of infinite discontinuity do not need to be considered one by one. For each function $g(x)$ and each positive integer N let $g_N(x)$ denote the lesser of $g(x)$ and N. If each $g_N(x)$ is Lebesgue-integrable in the sense already defined, expression (82) forms a nondecreasing sequence, and

$$\int_a^b g_N(x)\, dx \qquad (82)$$

so if expression (82) is bounded, it tends to a finite limit, which is taken as the value of integral (83). An

$$\int_a^b g(x)\, dx \qquad (83)$$

arbitrary function $g(x)$ is the difference of two nonnegative functions, Eqs. (84) and (85), and one may write Eq. (86) whenever both terms on the right have

$$g^+(x) = \frac{|g(x)| + g(x)}{2} \qquad (84)$$

$$g^-(x) = \frac{|g(x)| - g(x)}{2} \qquad (85)$$

$$\int_a^b g(x)\, dx = \int_a^b g^+(x)\, dx - \int_a^b g^-(x)\, dx \qquad (86)$$

a meaning according to the definition just given. It is provable that the class of all functions for which integrals exist according to the definitions just given does indeed have the property of completeness with respect to the Cauchy condition (79). The theory of Lebesgue yields many other useful results, including an extension of the fundamental theorem of integral calculus. One important restriction on the class of Lebesgue-integrable functions which is implicit in the definition is that when $g(x)$ is integrable so is $|g(x)|$. This restriction does not hold for the improper integrals which were described earlier.

The various methods of defining Lebesgue integrals are extensible also to functions of several variables. This extension throws light on the properties of multiple Riemann integrals and on the reduction of multiple integrals to repeated integrals.

Other definitions of integration. If $\alpha(x)$ is a fixed function defined on the interval $[a, b]$, the sum S defined by Eq. (1) may be replaced by that in Eq. (87).

$$S = \sum_{i=1}^{n} f(x_i)[\alpha(t_i) - \alpha(t_{i-1})] \qquad (87)$$

Then if the limit I of S exists in this case, it is called the Stieltjes integral of f with respect to α, and is denoted by integral (88). It has many of the properties of the Riemann integral, especially in case the function α is nondecreasing. In addition, it may take special account of the values of f at a finite or countable set of points. For example, if $\alpha(x) = c_i$ on $u_{i-1} < x < u_i$, where $u_0 < a$, $u_n > b$, $a < u_i < b$ for $i = 1, \ldots, n - 1$, and if $f(x)$ is continuous at each u_i, then Eq. (89) may be written. With some restrictions to ensure convergence, this may be extended to the case in which α has infinitely many discontinuities.

$$\int_a^b f(x)\, d\alpha(x) \qquad (88)$$

$$\int_a^b f(x)\, d\alpha(x) = \sum_{i=1}^{n-1} f(u_i)\,[c_{i+1} - c_i] \qquad (89)$$

In the case in which α is nondecreasing, the Stieltjes integral has an extension which is similar to that of the Lebesgue integral and is called the Lebesgue-Stieltjes integral.

Other extensions of the Lebesgue integral are the integrals of Denjoy. They apply to certain functions f for which $|f|$ is not integrable, and include the improper integrals, but not the Cauchy principal value. Numerous other types of integrals have been defined. In particular the integral of Lebesgue has been extended to cases in which the independent variable lies in a suitable space of infinitely many dimensions, or in which the functional values of f lie in such a space. *See* CALCULUS; FOURIER SERIES AND TRANSFORMS; SERIES.

Lawrence M. Graves

Bibliography. S.-B. Chae, *Lebesque Integration*, 2d ed., 1995; H. B. Dwight, *Tables of Integrals and Other Mathematical Data*, 4th ed., 1961; J. S. Gradshteyn and J. M. Ryzhjk, *Table of Integrals, Series, and Products*, 6th ed., 2000; P. R. Halmos, *Measure Theory*, 1950, reprint 1991; J. L. Kelley and T. P. Srinivasan, *Measure and Integral*, 1987; E. J. McShane, *Unified Integration*, 1983; W. Rudin, *Real and Complex Analysis*, 1987; C. W. Schwartz, *Measure, Integration and Function Spaces*, 1994; D. W. Stroock, *A Concise Introduction to the Theory of Integration*, 3d ed., 1998.

Integumentary patterns

All the features of the skin and its appendages that are arranged in designs, both in humans and other animals. Examples are scales, hairs, and feathers; coloration; and epidermal ridges of the fingers, palms, and feet. In its common usage, the term applies to the configurations of epidermal ridges, collectively named dermatoglyphics. Dermatoglyphics are characteristic of primates. *See* EPIDERMAL RIDGES; FEATHER; HAIR; SCALE (ZOOLOGY).

The superficial ridges are associated with a specific inner organization of skin. Skin is composed of two chief layers, the epidermis on the outside and the dermis underlying it (**Fig. 1**). These two layers are mortised by pegs of dermis, a double row of pegs corresponding to each ridge; these pegs accordingly form a patterning like that of the ridges.

Ridge patterns. The patterning of ridges, including that of the epidermal-dermal mortising, is determined during the third and fourth fetal months. All characteristics of single ridges and of their alignments are then determined definitively. Ridge alignments reflect directions of stress in growth of the hand and foot at the critical period of ridge differentiation. An important element in the production of localized patterns, for example, on the terminal segment of each digit, is the development in the fetus of a series of elevations, the volar pads.

Volar pads. The pads are homologs of the prominent pads on the paws of some mammals, but in primates they attain little elevation and soon tend to subside. The volar pads are disposed in a consistent topographic plan. Localized patterns have the same placement because growth of the pad is the determiner of the specific local pattern. When a pad has subsided before ridges are formed, its area does not present a localized pattern, and the ridges follow essentially straight, parallel courses. Variations in contours of the pads are accompanied by wide variations in the designs formed by the ridges overlying them.

Pattern variability. Variability of patterning is a major feature of dermatoglyphics and the basis for

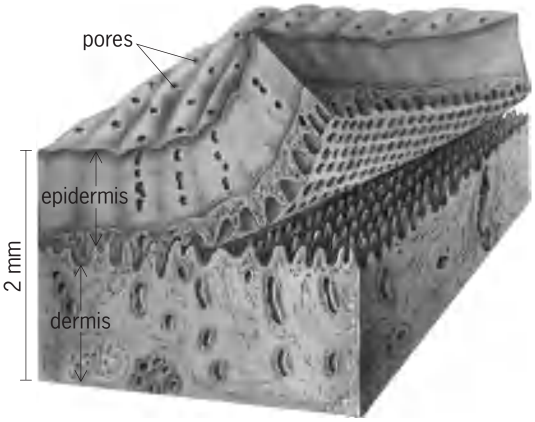


Fig. 1. Structure of ridged skin showing the chief two layers. (*From H. Cummins and C. Midlo, Finger Prints, Palms and Soles, McGraw-Hill, 1943*)

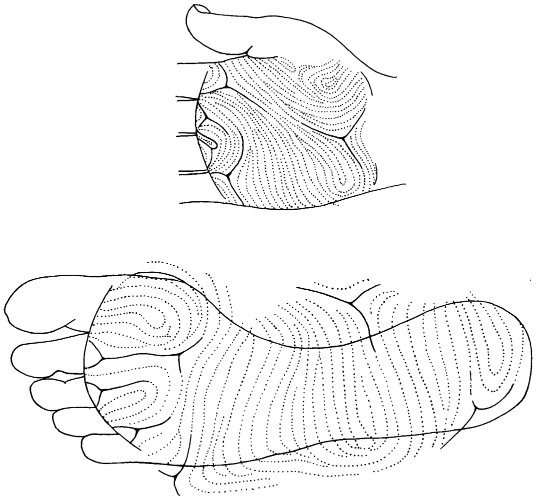


Fig. 2. Dermatoglyphics of palm and sole. (*After H. Cummins and C. Midlo*, *Finger Prints, Palms and Soles*, *McGraw-Hill, 1943*)

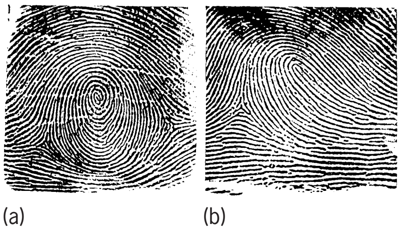


Fig. 3. Fingerprints. (*a*) Whorl pattern. (*b*) Loop pattern. (*From H. Cummins and C. Midlo, Finger Prints, Palms and Soles, McGraw-Hill, 1943*)

various applications (**Fig. 2**). In personal identification, prints (**Fig. 3**), customarily of fingers, are classified for filing in accordance with variables of pattern type and counts of ridges. Systematic filing makes it possible to locate readily and compare the sets of prints corresponding in classification to a set for which an identification is sought. In anthropological and medical investigations, groups of individuals are compared statistically in reference to the occurrence of these variables. Deductions may be drawn in accordance with likeness or unlikeness in the directions of variation. A few examples are cited. Trends of inheritance have been demonstrated in family groups and in comparisons of the two types of twins, fraternal (two-egg) and identical (one-egg). Dermatoglyphics thus are useful in diagnosing the types of twins and in analyzing cases of questioned paternity. Among different racial groups, the similar or discrepant trends of variation have been used to analyze racial affinities. Trends of variation are unlike in right and left hands, and the fact that they differ in accordance with functional handedness indicates an inborn predisposition of handedness. Departures from normal trends occur in many conditions associated with chromosomal aberrations, such as Down syndrome. *See* FINGERPRINT; HUMAN GENETICS; SKIN.

Harold Cummins

Bibliography. M. S. Elbauly and J. D. Schindler, *Handbook of Clinical Dermatoglyphs*, 1971; R. D. Olsen, Sr., *Scott's Fingerprint Mechanics*, 1978; R. I. Spearman, *The Integument: A Textbook of Skin Biology*, 1973.

Intelligence

General mental ability due to the integrative and adaptive functions of the brain that permit complex, unstereotyped, purposive responses to novel or changing situations, involving discrimination, generalization, learning, concept formation, inference, mental manipulation of memories, images, words and abstract symbols, eduction of relations and correlates, reasoning, and problem solving.

Theory. Verbal definitions or descriptions of intelligence, of examples of intelligent behavior, or of the characteristics of tests that best measure intelligence are all inadequate for a scientific understanding of intelligence. This is because “intelligence” is not a denotative noun, but a hypothetical construct or theory that is needed to account for one of the most striking and best-established phenomena in all of psychology: the ubiquitous positive intercorrelations among measurements of individual differences on all tests of “mental ability,” however diverse. By “mental” is meant only that individual differences in the test scores are not the result of individual differences in sensory acuity or motor dexterity and coordination. By “ability” is meant conscious, voluntary performance on a task (for example, a test item) for which it can be reliably determined whether or not the performance meets some objective standard of correctness or efficiency.

The British psychologist Charles E. Spearman was the first to discover that all mental tests show positive correlations with one another when the scores are obtained in a representative sample of the general population. From this he inferred that all mental tests of every kind must measure some factor of ability that they all measure in common. He termed this common factor “general mental ability” and gave it the label *g*. Spearman, and most psychologists following him, have identified “general intelligence” with *g*.

By means of factor analysis, a quantitative method developed by Spearman for analyzing a matrix of all of the intercorrelations among a number of tests, it became possible to determine the proportion of the total variation in all of the test scores that can be attributed to variation in the general factor *g* common to all of the tests, and to determine the degree to which each test measured the *g* factor. A test's *g* loading, as derived from factor analysis, can be thought of as the test's coefficient of correlation with the *g* factor. Factor analyses of more than a hundred varieties of tests or test items reveal a wide range of *g* loadings. By comparing the tests having high, moderate, and low *g* loadings, one can gain an inkling of the nature of *g*. It cannot be identified with any specific type of informational content or skill required by the test. In these respects, highly *g*-loaded items are as diverse as all items in general. For example, tests of vocabulary, problem arithmetic, and block designs (copying designs with multicolored blocks) are all highly *g*-loaded (and by their very nature are highly correlated with one another), whereas tests of spelling ability and arithmetic computation are only moderately *g*-loaded. Spearman, after inspecting hundreds of types of tests in relation to their *g* loadings, concluded that the only common features of the most highly *g*-loaded tests are that they involve "the eduction of relations and correlates" and "abstractness."

In more general terms, it is found that *g* is related to the complexity of the mental operations that are required by the task. Tasks involving reasoning and problem solving are much more *g*-loaded than tasks involving rote learning, routine skills, or the recall of specific information. When diverse tests are rank-ordered by their *g* loadings, they fall into four groups, from highest to lowest *g* loadings; the test items in these groups can be characterized as relational, associative, perceptual, and sensory-motor.

As yet there is no satisfactory or generally accepted theory of the neural mechanisms responsible for *g* or for individual differences in *g*, although there has been no paucity of speculation involving such notions as the total neural energy available to the brain, the number of brain cells and synaptic connections, and the speed of neural conduction.

The sexes, on average, do not differ in *g*; however, females slightly excel in verbal ability, and males slightly excel in spatial and quantitative abilities.

Factor analysis has revealed about 30 other factors of mental ability besides *g*. The most important of these factors, in terms of the amount of individual variation they account for, are verbal ability, numerical ability, spatial-visualization ability, and memory. A number of persons all with the same level of general intelligence or *g* may differ markedly from one another in these other ability factors. *See* FACTOR ANALYSIS.

Unity or modularity of intelligence. Some of the highly correlated abilities identified as factors probably represent what are also referred to as modules. Modules are distinct, innate brain structures that have developed in the course of human evolution. They are especially characterized by the various ways that information or knowledge is represented by the neural activity of the brain. Thus, the main modules are linguistic (verbal, auditory, lexical, and semantic), visuo-spatial, object recognition, numerical-mathematical, musical, and kinesthetic.

Although modules generally exist in all normal individuals, they are most strikingly in those with highly localized brain damage or pathology, and idiot savants. Savants evince striking discrepancies between a particular narrow ability and nearly all other abilities, often showing the greatest inconsistency with the individual's low level of general ability. There are some savants who are too mentally retarded to take care of themselves, yet can perform feats of mental calculation, or play the piano by ear, or memorize pages of a telephone directory, or draw objects from memory with nearly photographic accuracy. The modularity of these abilities is evinced by the fact that rarely, if ever, is more than one of them seen in a given savant. In contrast, there are individuals whose tested general level of ability is within the normal range, yet who, because of a localized brain lesion, show a severe deficiency in some particular ability, such as face recognition; receptive or expressive language dysfunctions (aphasia); or inability to form long-term memories of events. Again, modularity is evinced by the fact that these functional deficiencies are quite isolated from the individual's, total repertoire of abilities. In individual's with a normally intact brain, a module's efficiency can be narrowly enhanced through extensive experience and practice in the particular domain served by the module.

Measurement. Intelligence tests are diverse collections of *g*-loaded tasks (or items), graded in difficulty. The person's performance on each item can be objectively scored (for example, pass or fail); the total number of items passed is called the raw score. Raw scores are converted to some form of scaled scores which can be given a statistical interpretation.

The first practical intelligence test for children, devised in 1905 by the French psychologist Alfred Binet, converted raw scores to a scale of "mental age," defined as the raw score obtained by the average of all children of a given age. Mental age (MA) divided by chronological age (CA) yields the well-known intelligence quotient or IQ. When multiplied by 100 (to get rid of the decimal), the average IQ at every age is therefore 100, with a standard deviation of approximately 15 or 16. Because raw scores on mental tests increase linearly with age only up to about 16 years, the conversion of raw scores to a mental-age scale beyond age 16 must resort to statistical artifices. Because of this problem and the difficulty of constructing mental-age scales which preserve exactly the same standard deviation of IQs at every age, all modern tests have abandoned the mental-age concept and the calculation of IQ from the ratio of MA to CA. Nowadays the IQ is simply a standardized score with a population mean of 100 and a standard deviation (σ) of 15 at every age from early childhood into adulthood. The population distribution of IQ conforms closely to the normal curve in the range of $\pm 2\sigma$ from the mean (that is, from IQ 70 to 130), but there are excesses beyond that range. The middle 50% considered "average," fall between IQs of 90 and 110. IQs below 70 generally indicate "mental retardation," and above 130, "giftedness." The vast majority of students in selective colleges range above IQ 110, and graduates of 4-year colleges average about 120–125. *See* MENTAL RETARDATION.

Correlates of IQ. The IQ, as a measure of *g*, gains its importance from its substantial correlations with many educationally, occupationally, and socially valued variables. No other single measurement that can be obtained on a child (including parents' education and socioeconomic status) better predicts the child's scholastic performance. Childhood IQ, because of its relationship to later educational attainment, also predicts adult occupational status.

The fact that the IQ correlates with physical variables such as brain size and the amplitude and speed of evoked electrical potentials of the brain indicates that IQ tests reflect something biological, and not merely knowledge or skills acquired in school or in a cultured home.

Genetics. Individual differences in IQ involve polygenic inheritance, much as other continuous characteristics such as height and weight. The proportion of the total phenotypic variance in the population attributable to genetic factors, termed heritability, can be estimated by the methods of quantitative genetics from the correlations between various kinships—identical and fraternal twins, siblings, parent-child, and adopted children. Virtually all of the methodologically defensible estimates of the heritability of IQ range between.50 and.80, with a median value of about.70. Thus genetic factors are about twice as important as environmental factors as a cause of individual variation in intelligence.

All levels of intelligence are found within every major racial group. There is no definitive scientific evidence or consensus on the relative contributions of genetic and environmental factors to any undisputed average differences in intelligence between races. *See* BEHAVIOR GENETICS. Arthur R. Jensen

Bibliography. M. Anderson, *Intelligence and Development: A Cognitive Theory*, 1992; N. Brody, *Intelligence*, 2d ed., 1997; J. B. Carroll, *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*, 1993; D. K. Detterman (ed.), *Current Topics in Human Intelligence*, vol. 2: *Is Mind Modular or Unitary?*, 1992; J. A. Fodor, *The Modularity of Mind*, 1983; R. J. Sternberg (ed.), *Encyclopedia of Intelligence*, 1994.

Intelligent machine

Any machine that can accomplish its specific task in the presence of uncertainty and variability in its environment. The machine's ability to monitor its environment and then adjust its actions based on what it has sensed is a prerequisite for intelligence. The term intelligent machine is an anthropomorphism in that intelligence is defined by the criterion that the actions would appear intelligent if a person were to do it. A precise, unambiguous, and commonly held definition of intelligence does not exist.

Examples of intelligent machines include industrial robots equipped with sensors, computers equipped with speech recognition and voice synthesis, self-guided vehicles relying on vision rather than on marked roadways, and so-called smart weapons, which are capable of target identification. These varied systems include three major subsystems: sensors, actuators, and control. The class of computer programs known as expert systems is included with intelligent machines, even though the sensory input and output functions are simply character-oriented communications. The complexity of control and the mimicking of human deductive and logic skills makes expert systems central in the realm of intelligent machines. *See* EXPERT SYSTEMS; GUIDANCE SYSTEMS; ROBOTICS; VOICE RESPONSE.

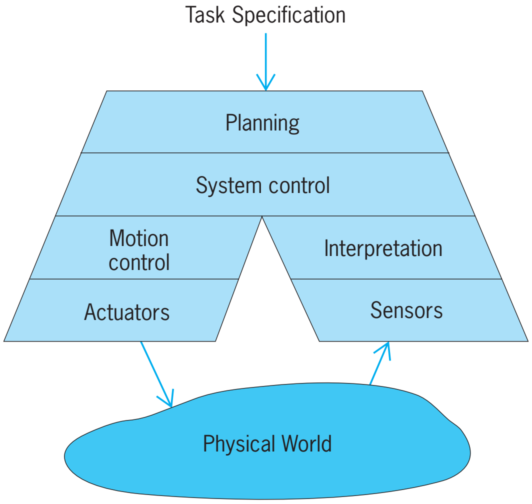


Fig. 1. Flow of information and data in a typical intelligent machine.

Intelligent control. Since the physical embodiment of the machine or the particular task performed by the machine does not mark it as intelligent, the appearance of intelligence must come from the nature of the control or decision-making process that the machine performs. Given the centrality of control to any form of intelligent machine, intelligent control is the essence of an intelligent machine. The control function accepts several kinds of data, including the specification for the task to be performed and the current state of the task from the sensors. The control function then computes the signals needed to accomplish the task. When the task is completed, this also must be recognized and the controller must signal the supervisor that it is ready for the next assignment (**Fig. 1**).

Automatic, feedback, or regulatory systems such as thermostats, automobile cruise controls, and photoelectric door openers are not considered intelligent machines. Several important concepts separate these simple feedback and control systems from intelligent control. While examples could be derived from any of the classes of intelligent machines, robots will be used here to illustrate five concepts that are typical of intelligent control. (1) An intelligent control system typically deals with many sources of information about its state and the state of its environment. A robot may contain a vision system and tactile sensors in addition to the internal position sensors that allow calculation of its spatial location. Given multiple sources of information, the intelligent control function evaluates them as needed based on the nature of the functions that must be performed at any instant. (2) An intelligent control system can accommodate incomplete or inconsistent information. For example, a robot expecting to pick up a single object from a conveyor belt may be confronted with two overlapping objects or no objects. (3) Intelligent control is characterized by the use of heuristic methods in addition to algorithmic control methods. A heuristic is a rule of thumb, a particular solution or strategy to be used for solving a problem that can be used for only very limited ranges of the input parameters. A possible heuristic for solving a particular robot-implemented insertion task would be to move the part incrementally along an outward spiral path, retrying the insertion, if the original attempt should fail. This heuristic could be repeated many times until an area large compared to positional uncertainties was covered. Clearly, such a heuristic is applicable only to a very small fraction of robot tasks. (4) An intelligent machine has a built-in knowledge base that it can use to deal with infrequent or unplanned events. An intelligent robot would have mechanisms, both sensors and heuristics, ensuring its own and its operator's safety. (5) An algorithmic control approach assumes that all relevant data for making decisions is available. The heuristic approach is predicated on the knowledge that all relevant data cannot be made available, even in principle, and that the control system will have to resolve ambiguous cases. *See* ADAPTIVE CONTROL; CONTROL SYSTEMS.

Intelligent robots. To clarify these very general concepts and to illustrate them with more detail, a sensor-equipped robot performing an electronics assembly task will be described. The task is to pick a part from an egg-carton-like pallet, inserting it into a test fixture, and, if it tests acceptably, to insert it into the appropriate place on the printed wiring board being assembled. A human worker, especially one with some experience in production line assembly, could be instructed in this task in a few minutes and would also know, without further instruction, what actions to take for the many exceptional conditions that can arise in this type of activity. Outlining the steps that must be taken to program a robot to accomplish the same task will illustrate the amount of prior knowledge possessed by a person and the general level of complexity of programming in an environment where all relevant data and actions must be stated completely and unambiguously. The robot program will be given in verbal outline form rather than coded in a particular robot programming language. The numbers are used both for specifying branches in the flow of control and for referencing in the discussions that follow.

1. Wait for a printed wiring board to enter the assembly station.
2. Take an image of the parts presentation pallet: an egg-carton arrangement that separates and loosely orients the parts.
3. Locate a part in the image.
4. If the pallet is empty, move it to the stacking area and signal the conveyor to move a fresh pallet into position. Resume at step 2.
5. If a part is present, pick up the part and position it in the visual inspection station. Take an image of the bottom of the part. Determine the location and orientation of the pin field.
6. If an incorrect pin field is seen, deposit the part in the reject bin and resume at step 2.
7. If the pin field is correct, orient the robot hand based on the observed orientation of the part and insert the part into the electrical test station.
8. Signal the electrical test to begin.
9. If the test fails, remove the part from the test fixture and deposit it in the reject bin. Then resume at step 2.
10. If the electrical test was passed, regrasp the part to ensure it is accurately held by the robot. Remove it from the test fixture and move it to a point just above the insertion location on the printed wiring board.
11. Insert the part in the printed wiring board by moving it vertically downward while monitoring the force. When the vertical force reaches a specified value, stop the robot's vertical motion.
12. Check the vertical position of the robot hand when it stopped moving. If it is at a height indicating that the part was inserted, release the part, retract the robot hand, and resume at step 1.
13. If the part was not inserted correctly, move the robot hand in the plane of the printed wiring board by a small increment along a prescribed path and resume at step 11.

Several comments can be made about a robot that is following the above program and the program itself. Such a robot would appear to be intelligent given the definition in the beginning of the article. There is variability in the robot's environment, and it is using a variety of sensory data to guide its actions. However, the program is incomplete. There are many things that can go wrong that have not been accounted for: If a hole in the printed wiring board for the part is missing, the insertion cannot be successfully accomplished. In this case, program steps 11, 12, and 13 form an apparent infinite loop. Many other similar problems lurk in this program. The program also presupposes significant capabilities and data in the robot and related vision system. The vision system and robot must be calibrated; that is, a procedure must be periodically executed that calculates the transformations between the robot and vision system coordinates. The vision system must be programmed to recognize first what the part looks like in the context of the pallet (step 3) and second what a correct pin field looks like (step 5). The precise locations of all points of reference for the robot must be determined. The search strategy to be followed if the part is not inserted correctly (step 13) must be carefully defined. The calibration and teaching programs required to accomplish these steps may be as complex as the primary task.

Execution of the assembly program and all of the implied subsidiary functions defines the control capability of the robot. The threshold of intelligent control is reached when the system can deal with many of the infrequent and exceptional conditions without requiring outside intervention or assistance. A second part of intelligent control is the development of the detailed programs given a high-level statement of the task requirements. This is the planning function (Fig. 1). Systems of this level of complexity are primarily found in robotics research laboratories. The automatic planning of complex robot control programs given high-level task descriptions is an area of research.

Control hierarchy. The control function in an intelligent machine is generally implemented as a hierarchy of processes. The upper levels are responsible for global decision-making processes and for the planning phases of the task. The lower levels implement critical time-dependent subtasks. In **Fig. 2** the interconnection of a group of processors forming a vision-controlled robot is diagrammed. The manipulator arm is itself controlled by a two-level hierarchy. The lower level consists of a processor that performs the servo function for each joint of the arm. The robot controller converts world coordinates supplied by the system controller into the values for each joint required to place the robot hand in the desired location. The robot controller also divides complete movement specifications into a series of short steps that result in coordinated motion and controlled velocity of the robot. *See* SERVOMECHANISM.

The control program outlined in the robot example would be executed on the system-level control processor. In addition to the vision, manipulator, and safety subsystems, the system-level controller would also be the interface with the robot operator and the factory communications system. The partitioning of subtasks onto autonomous communicating processors is a natural and economic way to obtain increased system performance by using multiple microprocessors. *See* MICROPROCESSOR.

Computer vision. A computer vision system interprets the contents of a large array (256 × 256 is common) of intensity values that is the computer representation of a frame of a television image. Interpret in this context means to find, recognize, or inspect a prespecified object or group of objects. The first processing step is to isolate or segment the image of the object from the background image. The most common technique in use is to create a high-contrast image, typically by backlighting the object of interest, which allows a simple threshold to be used for the segmentation. Objects in the image are found at those locations where the intensity value is less than some prespecified constant. Lower-contrast images, such as a part in a pallet, require segmentation techniques that rely on rapid changes in brightness (edges) or differences in intensity between two images, or techniques that look for specific visual features by using any of a large number of matched filter techniques. The amount of computation required to segment a low-contrast image is significantly greater than that required by thresholding. Once an object image is segmented from the background, a set of parameters known as features is calculated. The set of features, which may include the object area, perimeter, number of holes and any other enumerable parameter, is compared to a model of the object either to ascertain its identity or to compare it to some inspection criteria. The specific set of features needed for a specific visual recognition problem depends on the exact appearances of the objects in complex ways with the result that the feature set is normally determined empirically. Similarly, orientation and location quantities are also computed for the segmented image. Generally, an image represents a two-dimensional projection of a three-dimensional object. Correction to the perceived shape and location may have to be made based on its location relative to the viewing camera. No matter what complexity of techniques are used in its construction, a computer vision system is invariably implemented with a dedicated processor. *See* COMPUTER VISION.

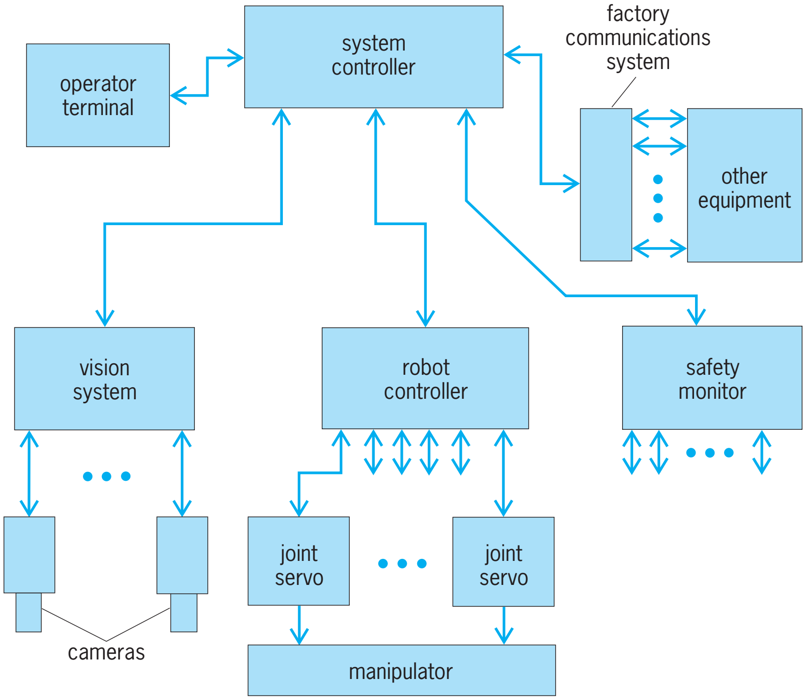


Fig. 2. **Interconnection of processors forming a sensor-based robot.**

Interconnection of machines. Another trend expected in the area of intelligent machines is their interconnection into cooperating groups to address larger tasks. Standardization of information exchange protocols and languages describing various fields of work are prerequisites to any extensive development of interconnected, cooperating systems.

Capabilities. Slow, steady improvements may be expected in intelligent machine capabilities. Progress will result from advances in microprocessors and the implementation of various functions as very large-scale integrated (VLSI) circuits. *See* INTEGRATED CIRCUITS.

Expert systems. The control schemes outlined above are algorithmic even if the resulting system behavior appears intelligent. The explicit programming methods described do not lend themselves to situations where several alternative actions could be rationally considered. In the explicit programming of a complex task, there is very little opportunity to use default actions for many infrequent conditions. The method of expert systems is expected to provide a paradigm for a better way of robot task planning and execution. Expert systems have a knowledge base and an inference procedure for obtaining a result consistent with the contents of the knowledge base, given an arbitrary set of input conditions. The knowledge base is represented by a collection of condition-action statements. These statements are called rules and the inference procedure is known as the rule interpreter. The other major components of an expert system (discussed below) are a user interface, a database, and a rule structure.

The primary difficulties are in obtaining an adequate collection of rules for the task and in defining an appropriate strategy for resolving inconsistencies or ambiguities when applying rules to a given problem. A specific robot task would be programmed by adding a set of rules, describing the particular operation to be formed, to an existing, and probably much larger, set of rules that defined the default and exception-handling behavior of the robot. John F. Jarvis

Applications. Since the mid-1970s, expert systems have been continuously developed and applied by industry, business, and commerce. Expert systems are the most successful implementation of artificial intelligence. Applications include medical diagnosis, mineral analysis, control of complex processes, and fault diagnosis and monitoring; systems for fault diagnosis and monitoring account for about half the applications. *See* ARTIFICIAL INTELLIGENCE.

Goal-directed systems. In many expert systems, the rule interpreter, or inference engine as it is sometimes called, uses an exhaustive backward-chaining strategy, in which the rules are expressed in a highly focused "IF-condition-THEN-result" form, the result is known or concluded, and, knowing the answer, the system searches backward for the rule whose condition produced the result. Such an expert system is termed goal-directed, since the search is, in fact, directed toward a goal (the known result). Goal-directed systems are therefore the result of going from the general to the particular.

An inference mechanism based on a backward-chaining strategy is commonly equipped with an extensive explanation facility to reduce the amount of conflict resolution needed. To be user-friendly, an expert system must also possess a user interface that interrogates the human expert in a manner that is unambiguous. The information obtained from the user interface must then be stored as facts in a database, and the relationship between individual facts must be stated. The rule structure determines which rule should be examined next, in a forward-chaining expert system, or it further questions the user in a backward-chaining system. Finally, the representational structure must also include an uncertainty-handling mechanism that is based on measurements of belief. *See* COMPUTER; DATABASE MANAGEMENT SYSTEM; FUZZY SETS AND SYSTEMS.

While the use of rules having the form "IF-condition-THEN-result," known as production rules, has been particularly successful when used in goal-directed applications such as diagnostics, such rules have limitations. The major limitations associated with production rules are the so-called knowledge acquisition bottleneck and the difficulty of determining the correct method of representing the rules when there is a large number of them. In order to overcome these limitations, alternative representations have been attempted. Production rules have been represented in the form of inference nets, causal mechanisms, and frame-based abduction mechanisms. The inference net is overlaid by a semantic net to connect different objects. As with production rules, an inference-net approach can be backward-chained, but it requires a different mechanism for handling uncertainty. For inference nets, a mechanism based on Bayes' rule, a statistical rule, has been used. However, if an expert system is constructed by using causal or frame-based abduction mechanisms it must have inference and control mechanisms suited to those representations. Also, the relational mechanism for storing facts in the database must alter accordingly. For causal nets the database is divided into findings and hypotheses. In frame-based abduction the database holds domain knowledge in the form of descriptive frames of information. Also, with frame-based abduction the inference mechanism uses hypotheses-and-test cycles. *See* ESTIMATION THEORY.

Machine learning. The difficulties surrounding production rules, including knowledge acquisition, descriptive knowledge representation, and context handling, have led to the development of a subdivision of expert system technology known as machine learning, which is a method for automatically generating rules from examples. Thus, machine learning is an approach based on the idea of going from the particular to the general.

Machine learning can be classified into three areas, artificial-intelligence–type learning, neural nets, and genetic algorithms. An effective machine-learning system must use sampled data to generate internal updates, be capable of explaining its findings in an understandable way (for example, symbolically), and communicate its explanation to a human expert in a manner that improves the human's understanding. Expert systems based on artificial-intelligence learning have surpassed the performance of human experts, and they have been able to communicate what they have learned to human experts for the purpose of verification.

There are two schools of artificial-intelligence-type learning. One is based on logic and attempts to find minimal representations of finite functions, while the other concentrates on learning via knowledge representation, acquisition, compaction, and learning from example. The theoretical base of machine learning is multivariate analysis in statistics, and the use of decision trees. Attribute-based learning systems use a common form of knowledge representation to learn explicit rules. Input attributes can be discrete, continuous, or tree-structured, or they can be example sets that contain attribute values. In example sets, each distinguished attribute is the classification attribute for each example. The output is a decision rule that is capable of classifying further, unseen examples that are presented to the system. *See* DECISION THEORY; STATISTICS.

Artificial-intelligence–type learning originated from an investigation into the possibility of using decision trees or production rules for concept representation. Subsequently the work was extended to use decision trees and production rules in order to handle the most conventional data types, including those with noisy data, and to function as a knowledge acquisition tool. To construct a rule from a training set, the algorithm must first examine all the examples in the training set to see if they are all of the same class. If this is the case, classification is achieved. If the training sets are not all of the same class, then the most significant attribute is selected from the training set, the set is partitioned into two sets, and subtrees are constructed. The evaluation function used for selecting an attribute is based on information theory, and recursive methods are used to construct the subtrees. Based on this approach, in 1978 the first expert system generated exclusively by inductive learning from example was used to diagnose 19 soybean diseases with a degree of accuracy higher than that of a human expert.

Interest in neural-net (connectionist) learning algorithms stems from their ability to construct a network model based on training examples. When finally constructed, the neural network becomes the knowledge base for an expert system. Neural networks are capable of automatically generating behavior based on example sets of types described in the discussion of artificial-intelligence–type learning above. Their major limitation is that they are useful only in diagnostic systems that have fixed input and output sets. *See* NEURAL NETWORK.

Genetic algorithms are based on iteration. They use general-purpose search strategies that have their foundations in genetics, in particular, in the process of natural selection. The purpose of iteration is to weed out weaker representations and give strengthened support to those that provide better results. Genetic algorithms attempt to balance exploration with exploitation, yet at the same time they try to avoid the problems associated with other goal-directed methods. For example, they attempt to avoid the local-minima problems found in hill climbing methods and the inefficiencies of random search. Since they make no prior assumption about the problem domain, they can be applied generally, and, in particular, in areas whose characteristics are unknown. They have been applied successfully in domains where optimization is a problem, but they started to be used for learning systems only in the late 1980s. The iterative procedure of a genetic algorithm maintains a set of knowledge structures, called a population. This structure is a string of symbols chosen from a fixed set of examples; this is a candidate solution. At each generation the structure is evaluated and is given a fitness value that is dependent on how close it is to the required solution. While genetic algorithms and neural nets have contributed to the study of machine learning, a disadvantage of these methods is that learning is a slow process when they are used. *See* EXPERT SYSTEMS.

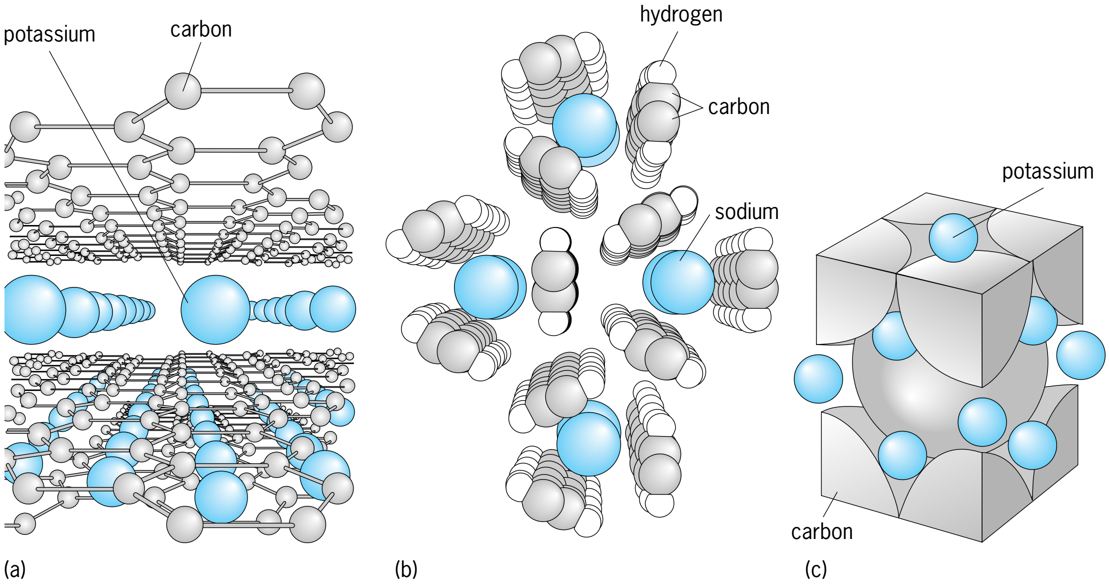
Edward Grant

Bibliography. D. A. Bradley et al., *Mechatronics and the Design of Intelligent Mechanics and Systems*, 2000; R. Kurzweil, *Age of Intelligent Machines*, 1992; M. W. Spong, S. Hutchinson, and M. Vidyasagar, *Robot Modeling and Control*, 2005; G. Zobrist and C. Y. Ho, *Intelligent Systems and Robotics*, 2000.

Intercalation compounds

Crystalline or partially crystalline solids consisting of a host lattice containing voids into which guest atoms or molecules are inserted. Candidate hosts for intercalation reactions may be classified by the number of directions (0 to 3) along which the lattice is strongly bonded and thus unaffected by the intercalation reaction. Isotropic, three-dimensional lattices (including many oxides and zeolites) contain large voids that can accept multiple guest atoms or molecules. Layer-type, two-dimensional lattices (graphite and clays) swell up perpendicular to the layers when the guest atoms enter (**illus.** *a*). The chains in one-dimensional structures (polymers such as polyacetylene) rotate cooperatively about their axes during the intercalation reaction to form channels that are occupied by the guest atoms (illus. *b*). In the intercalation family based on solid C_{60} (buckminsterfullerene), the zero-dimensional host lattice consists of 60-atom carbon clusters with strong internal bonding but weak intercluster bonding. These clusters pack together like hard 1-nm-diameter spheres, creating interstitial voids which are large enough to accept most elements in the periodic table (illus. *c*). The proportions of guest and host atoms may be varied continuously in many of these materials, which are therefore not true compounds. Many ternary and quaternary substances, containing two or three distinct guest species, are known. The guest may be an atom or inorganic molecule (such as an alkali metal, halogen, or metal halide), an organic molecule (for example, an aromatic such as benzene, pyridine, or ammonia), or both. *See* CRYSTAL STRUCTURE; FULLERENE; GRAPHITE; POLYMER; ZEOLITE.

Competing interactions. The stability of a particular guest-host configuration depends on the balance between positive and negative contributions to the formation energy of the so-called compound. For example, in carbon-based hosts such as graphite, both guest and host become ionized after intercalation, and the attraction between oppositely charged guest and host ions helps to offset the elastic energy penalty associated with the swelling of the lattice. Some of these contributions vary with temperature, leading to structural phase transitions such as internal melting of the guest layers in graphite intercalation compounds. In some cases the guest layers



Typical intercalation compounds. (*a*) Potassium in graphite, KC_8, a prototype layer intercalate. (*b*) Sodium in polyacetylene, $[Na_{0.13}(CH)]_x$, where *x* denotes infinitely repeating polymer chains. (*c*) Potassium in solid buckminsterfullerene, K_3C_{60}. The internal structure of the fullerene molecule is not shown. (*After J. E. Fischer and P. Bernier, Les cristayx des fullerènes, La Recherche, 24(250):46–55, January 1993*)

can be far enough apart to be considered independent, providing important prototypes for studying phenomena such as magnetism, superconductivity, and phase transitions in two dimensions. The spatial distribution of guest atoms or molecules within the host network can be regular or disordered, depending on temperature, concentration, and the chemical nature of the guest. *See* COHESION (PHYSICS); SOLID-STATE CHEMISTRY.

An example of temperature-dependent behavior is exhibited by alkali atoms inserted in graphite. If precisely half the number of alkali atoms that could be accommodated at the limit of close-packed alkali monolayers is added between each pair of graphite sheets, the system displays a phase transition. At low temperature, the elastic penalty will be minimized if the alkali ions are crowded into only half the galleries, resulting in a periodic dense-stage-2 configuration: maximally dense alkali monolayers between every other pair of host sheets. At high temperature the mixing entropy between filled and empty sites drives a transition to a site-disordered dilute-stage-1 configuration, in which all galleries contain some ions, on average 50% of the hard-sphere limit. The dense-stage-2 configuration may be recovered by applying hydrostatic pressure to the high-temperature configuration, because the volume is minimized by segregating the filled and empty sites into completely filled and empty galleries respectively. *See* ENTROPY; PHASE TRANSITIONS; THERMODYNAMIC PRINCIPLES.

Properties. The physical properties of the host are often dramatically altered by intercalation. In carbon-based hosts such as graphite, polyacetylene, and solid C_{60}, the charge donated to the host by the guest is delocalized in at least one direction, leading to large enhancements in electrical conductivity. Some hosts can be intercalated by electron-donating or -accepting guests, leading respectively to *n*-type or *p*-type synthetic metals, analogous to doped semiconductors. The chemical bonding between guest and host can be exploited to fine tune the chemical reactivity of the guest. In many cases, superconductors can be created out of nonsuperconducting constituents. *See* SEMICONDUCTOR; SUPERCONDUCTIVITY.

Applications. Many applications of intercalation compounds derive from the reversibility of the intercalation reaction. The best-known example is pottery: Water intercalated between the silicate sheets makes wet clay plastic, while driving the water out during firing results in a dense, hard, durable material. Many intercalation compounds are good ionic conductors and are thus useful as electrodes in batteries and fuel cells. A technology for lightweight rechargeable batteries employs lithium ions which shuttle back and forth between two different intercalation electrodes as the battery is charged and discharged: vanadium oxide (three-dimensional) and graphite (two-dimensional). Zeolites containing metal atoms remain sufficiently porous to serve as catalysts for gas-phase reactions. Many compounds can be used as convenient storage media, releasing the guest molecules in a controlled manner by mild heating. *See* BATTERY; CLAY, COMMERCIAL; FUEL CELL; SOLID-STATE BATTERY.

J. E. Fischer

Bibliography. P. Bernier et al. (eds.), *Chemical Physics of Intercalation II*, 1993; T. Enoki, M. Suzuki, and M. Endo, *Graphite Intercalation Compounds and Applications*, 2003; C. Julien, J. P. Pereira-Ramos, and A. Momchilov (eds.), *New Trends in Intercalation Compounds for Energy Storage*, 2002; A. P. Legrand and S. Flandrois (eds.), *Chemical Physics of Intercalation*, 1987.

Intercommunicating systems (intercoms)

A privately owned system that allows voice communication between a limited number of locations, usually within a relatively small area, such as a building, office, or residence. Intercommunicating systems are generally known as intercoms. Intercom systems can vary widely in complexity, features, and technology. Though limited in size and scope, intercom systems can provide easy and reliable communication for their users.

Wire systems. An extremely simple intercom is a two-station arrangement in which one station is connected to the other via a dedicated wire. Simple twisted-pair cabling is the most common type of wiring for this type of intercom. An example in a business application is an executive-secretary system, in which the executive presses a button to signal the secretary and initiate the connection. When the secretary answers, the voice link is automatically established.

Other systems have multiple stations, as many as 10 to 20, any of which can connect with any other station. The user must dial a one- or two-digit code to signal the intended destination. When the called party answers, the talk path is established between the two stations. In a residence, this type of system permits calling between extension telephones through abbreviated dialing codes; the called extension alerts the user through a distinctive ringing signal.

Telephone intercoms. Still other intercom systems work in conjunction with key and hybrid key telephone/private branch exchange telephone systems. They support internal station-to-station calling rather than access to outside lines. Many intercoms possess some or all of the features of that particular telephone system, such as call hold, call pickup, and conferencing. *See* KEY TELEPHONE SYSTEM; PRIVATE BRANCH EXCHANGE.

Normally the telephone intercom is incorporated in the same telephone instrument that is used to access the public switched network. Two primary types of intercom systems exist: the single-link and the multilink. The single-link system provides a single dedicated intercom talk path in the telephone system, no matter how many stations might utilize the intercom system. The multilink intercom system provides two or more dedicated intercom paths that allow more than one simultaneous intercom conversation to be in progress. *See* SWITCHING SYSTEMS (COMMUNICATIONS); TELEPHONE SERVICE.

Wireless intercoms. A third type of intercom is the wireless intercom system for intrabuilding communications, which consists of a base unit radio transmitter, equipped with an antenna, and a number of roving units tuned to different frequencies. The base can selectively communicate with individual roving units by dialing the code corresponding to each roving unit's specific frequency. Commercial systems usually have a range of 1–2 mi (2–3 km). Structural components of the building, such as steel or concrete, can interfere with transmission quality and limit effective range. *See* MOBILE COMMUNICATIONS.

Byron W. Battles; Vincent F. Rafferty

Bibliography. The Aries Group·MPSG, *Glossary of Terminology: Key Telephone Systems and Hybrid KTS/PBX Systems*, 1990; Center for Communications Management, Inc., *Telecommunications Dictionary and Fact Book*, 1984; G. Langley, *Telephony's Dictionary*, 2d ed., 1986.

Interface of phases

The boundary between any two phases. Among the three phases, gas, liquid, and solid, five types of interfaces are possible: gas-liquid, gas-solid, liquid-liquid, liquid-solid, and solid-solid. The abrupt transition from one phase to another at these boundaries, even though subject to the kinetic effects of molecular motion, is statistically a surface only one or two molecules thick.

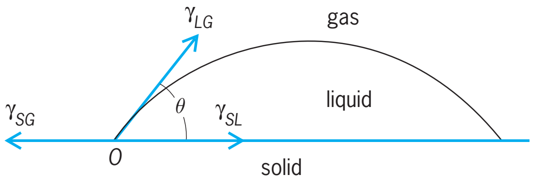
A unique property of the surfaces of the phases that adjoin at an interface is the surface energy which is the result of unbalanced molecular fields existing at the surfaces of the two phases. Within the bulk of a given phase, the intermolecular forces are uniform because each molecule enjoys a statistically homogeneous field produced by neighboring molecules of the same substance. Molecules in the surface of a phase, however, are bounded on one side by an entirely different environment, with the result that there are intermolecular forces that then tend to pull these surface molecules toward the bulk of the phase. A drop of water, as a result, tends to assume a spherical shape in order to reduce the surface area of the droplet to a minimum.

Surface energy. At an interface, there will be a difference in the tendencies for each phase to attract its own molecules. Consequently, there is always a minimum in the free energy of the surfaces at an interface, the net amount of which is called the interfacial energy. At the water-air interface, for example, the difference in molecular fields in the water and air surfaces accounts for the interfacial energy of 7.2×10^{-6} joule/cm^2 (11.15×10^{-6} cal/in.^{2}) of interfacial surface. The interfacial energy between the two liquids, benzene and water, is 3.5×10^{-6} J/cm^2 (5.4×10^{-6} cal/in.^{2}), and between ethyl ether and mercury is 37.9×10^{-6} J/cm^2 (58.67×10^{-6} cal/in.^{2}). These interfacial energies are also expressed as surface tension in units of millinewtons per meter.

The surface energy at an interface may be altered by the addition of solutes that migrate to the surface and modify the molecular forces there, or the surface energy may be changed by converting the planar interfacial boundary to a curved surface. Both the theoretical and practical implications of this change in surface energy are embodied in the Kelvin equation, Eq. (1) where P/P_0 is the ratio of the vapor pressure of

$$\ln \frac{P}{P_0} = \frac{2M\gamma}{RT\rho r} \qquad (1)$$

a liquid droplet with diameter r to the vapor pressure



Contact angle θ at interface of three phases.

of the pure liquid in bulk, ρ the density, γ the surface energy, and M the molecular weight. Thus, the smaller the droplet the greater the relative vapor pressure, and as a consequence, small droplets of liquid evaporate more rapidly than larger ones. The surface energy of solids is also a function of their size, and the Kelvin equation can be modified to describe the greater solubility of small particles compared to that of larger particles of the same solid. *See* ADSORPTION; SURFACTANT.

Contact angle. At liquid-solid interfaces, where the confluence of the two phases is usually termed wetting, a critical factor called the contact angle is involved. A drop of water placed on a paraffin surface, for example, retains a globular shape, whereas the same drop of water placed on a clean glass surface spreads out into a thin layer. In the first instance, the contact angle is practically 180°, and in the second instance, it is practically 0°. The study of contact angles reveals the interplay of interfacial energies at three boundaries. The **illustration** is a schematic representation of the cross section of a drop of liquid on a solid. There are solid-liquid, solid-gas, and liquid-gas interfaces that meet in a linear zone at O. The forces about O that determine the equilibrium contact angle are related to each other according to Eq. (2), where the γ terms represent free energies at

$$\gamma_{SG} = \gamma_{SL} + \gamma_{LG}\cos\theta \qquad (2)$$

the interfaces and θ is the contact angle. Since only γ_{LG} and θ can be measured readily, the term adhesion tension is defined by Eq. (3). Adhesion tension,

$$\gamma_{LG}\cos\theta = \gamma_{SG} - \gamma_{SL} = \text{adhesion tension} \qquad (3)$$

which is the free energy of setting, is of critical importance in detergency, dispersion of powders and pigments, lubrication, adhesion, and spreading processes.

The measurement of interfacial energies is made directly only upon liquid-gas and liquid-liquid interfaces. In measuring the liquid-gas interfacial energy (surface tension), the methods of capillary rise, drop weight on pendant drop, bubble pressure, sessile drops, Du Nuoy ring, vibrating jets, and ultrasonic action are among those used. There is a small but appreciable temperature effect upon surface tension, and this property is used to determine small differences in the surface tension of a liquid by placing the two ends of a liquid column in a capillary tube whose two ends are at different temperatures. The determination of interfacial energies at other types of interfaces can be inferred only by indirect methods. *See* FLOTATION; FOAM; FREE ENERGY; PHASE EQUILIBRIUM; SURFACE TENSION.

Wendell H. Slabaugh

Bibliography. A. W. Adamson and A. P. Gast, *Physical Chemistry of Surfaces*, 6th ed., 1997; H.-J. Butt, K. Graf, and M. Kappl, *Physics and Chemistry of Interfaces*, 2d ed., 2006.

Interference filters

Optical filters that are based on the interference of optical radiation in thin films—the phenomenon responsible for the iridescent colors observed in soap bubbles, oil slicks, shells, bird feathers, and butterfly wings. An optical filter is any material or device that is used to change the intensity, spectral distribution, phase, or polarization properties of the transmitted or reflected components of the optical radiation incident upon it. Large and complicated instruments, such as interferometers and spectrophotometers, are not usually included within this definition. Some of the many other different physical principles that have been used in the past to produce optical filters are absorption, reflection, scattering, diffraction, and polarization. The term "coating" is commonly used as a synonym for "filter" and, historically, has been applied to thin-film filters of certain functionalities. *See* COLOR FILTER; INTERFERENCE OF WAVES.

Filters and coatings. Filters and coatings based on thin-film interference are remarkably versatile. Although most are currently produced for the visible and adjacent parts of the spectrum, thin-film interference filters are also made for spectral regions extending from the x-ray to submillimeter wavelengths. Thus, the ratio of the highest to the lowest wavelengths for which optical interference coatings are made is of the order of $1{:}10^6$. *See* ELECTROMAGNETIC RADIATION.

In addition to their use in optical instruments for science and engineering, thin-film interference filters are used in eyeglasses and photographic cameras, energy conversion and conservation, architecture, communications, and display devices. New applications for optical interference filters appear regularly.

One disadvantage of filters based on thin-film interference is that unless special measures are taken, their properties change with the angle of incidence of light. In some cases, even that can be turned into an advantage. For example, it is possible to tune the filters by tilting them, and the iridescent property of the thin-film coatings can be used for protection from counterfeiting as well as for decorative purposes.

Thin-film thicknesses and materials. Optical thin films are mostly made of inorganic materials, including dielectrics (such as Al_2O_3, CaF_2, HfO_2, MgF_2, PbF_2, Sc_2O_3, SiO_2, TiO_2, Y_2O_3, ZnS, and ZrO), semiconductors (such as GaAs, Ge, and Si), metals (such as Ag, Al, Au, Cr, Ni, and Mo), and alloys (such as the Ni-Cr-Fe alloys, Inconel and Nichrome), although organic materials are increasingly being used. The thicknesses of the transparent films vary, but they are usually within one or two orders of magnitude of the wavelengths

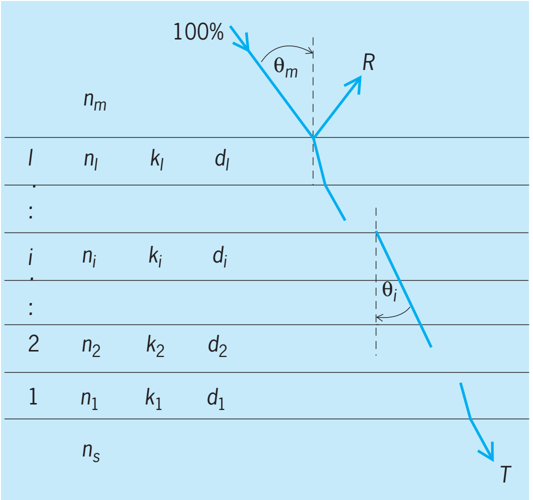


Fig. 1. Schematic representation of a homogeneous multilayer coating.

for which the filter is to be effective. The absolute thickness will vary a great deal, depending on whether the filter is intended for the x-ray, visible, or submillimeter spectral regions. For a filter designed for the visible part of the spectrum, a typical thickness might be around 100 nanometers. The thicknesses of metal and alloy layers within a multilayer system have to be thin enough that the layers are substantially transmitting. Metal layer thicknesses are typically around 5 nm.

Deposition processes. The most precise optical interference filters are currently produced by physical deposition processes, such as thermal or electron-beam evaporation with or without ion assist, by ion plating methods, by various forms of direct-current (dc), radio-frequency (rf), or alternating-current (ac) magnetron sputtering, or by ion-beam sputtering. All these methods require expensive vacuum equipment. Less precise coatings can be deposited by chemical vapor deposition or from a liquid phase.

Equipment for these processes is usually considerably cheaper, and is used for depositing thin-film filters for consumer products whenever the accuracy is adequate. When choosing the deposition process, other considerations are whether it lends itself for the manufacture of large areas of coatings with uniform thickness, whether the properties of the resulting coatings are chemically stable and do not change with time, whether the produced coating's electrical properties are damaged by the high power of laser systems, and whether the mechanical properties of the coatings will withstand the expected wear and tear. Meeting the nonoptical requirements is often more challenging than obtaining the required optical performance. *See* CHEMICAL VAPOR DEPOSITION; LASER; PHYSICAL VAPOR DEPOSITION; RADIATION DAMAGE TO MATERIALS; SPUTTERING.

Theory. In its simplest manifestation, a coating can consist of a single layer deposited on a glass surface. Some of the light incident on this film will be reflected from the air–film interface, some from the glass–film interface, and the remainder will be transmitted into the glass. Depending on the value of the refractive index of the coating material relative to that of the glass and the relative thickness of the film to the wavelength of the incident light, constructive or destructive interference will take place between the two reflected electromagnetic waves, and the reflectance of the glass–air interface onto which the layer has been deposited can be enhanced or reduced. For such a simple system, it is possible to write down explicit expressions for the resulting reflectance. But most applications require filters that consist not of one thin film but tens or hundreds of thin films. A schematic representation of an optical thin-film interference filter is shown in **Fig. 1**. The filter consists of l layers deposited onto a substrate of refractive index n_s and immersed in a medium of refractive index n_m. The parameters available to the thin-film designer to achieve the desired performance include n_s, n_m, and the refractive indices n_i, as well as extinction coefficients k_i, and thicknesses d_i of the l layers. Clearly, the more complicated the required spectral performance of a filter is, the greater the number of layers l and the higher the overall thickness of the multilayer coating that is needed to yield the required result. For accurate results during the design process, it is necessary to take into account not only the primary reflections from each interface but also the effects of multiply reflected beams from the various interfaces. The performance of such systems is best evaluated with the aid of computer programs that are based on Maxwell's equations for the propagation of electromagnetic waves. Depending on the design method used, the solution to the problem may consist of a single, thick inhomogeneous layer in which the optical constants vary gradually as a function of distance from the substrate, or of a multilayer system consisting of many homogeneous layers of two or more different materials. The latter are more readily implemented. *See* ABSORPTION OF ELECTROMAGNETIC RADIATION; MAXWELL'S EQUATIONS; OPTICAL MATERIALS; REFLECTION OF ELECTROMAGNETIC RADIATION; REFRACTION OF WAVES.

Optical interference filters and coatings. The following are examples of the most important types of optical interference filters and coatings.

Antireflection coatings. Antireflection coatings are the first application of optical thin-film interference. Around 1817, Joseph Fraunhofer discovered that tarnished glasses reflected less light than freshly polished glasses. He then proceeded to develop a process for artificially aging glasses to achieve this effect. The results were the first antireflection coatings. To this day, antireflection coatings are one of the most important optical thin-film components in both scientific and consumer instrumentation. The reflection of light from the surface of a transparent substrate at normal incidence can be as little as 4% for glass and as much as 30% for germanium, a material that is used extensively in infrared (IR) optics. Thus, germanium optics without antireflection coatings would lead to huge light losses. In

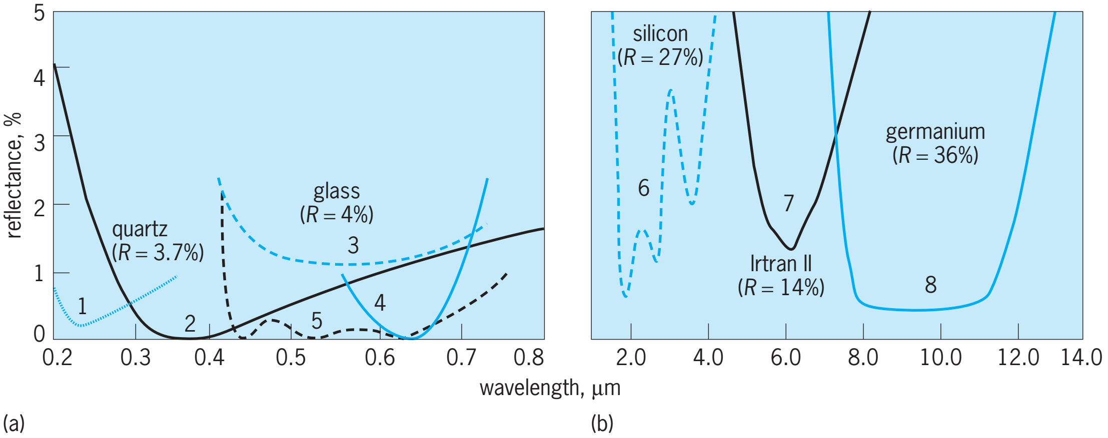


Fig. 2. Measured performance of antireflection coatings for quartz (curves 1, 2), glass (3–5), silicon (6), Irtran II [ZnS] (7), and germanium (8) surfaces in the (*a*) ultraviolet-visible and (*b*) infrared spectral regions. The numbers in brackets represent the reflectances of the surfaces without the antireflection coatings.

many instances, even glass surfaces must be antireflection-coated. For example, telephoto lenses for television and personal cameras can consist of many glass–air interfaces that must be antireflection-coated if the lenses are to be useful. This not only increases the light throughput of the objectives but also prevents the formation of stray light though multiple reflections that would reduce the contrast in the image plane. **Figure 2** shows the reflectances of substrate–air interfaces of various materials before and after treatment with antireflection coatings. Unless otherwise stated, the curves shown in Fig. 2 and subsequent diagrams are measured values. At the present state of thin-film deposition technology, the difference between theory and experiment is not very significant in most cases. Another special type of antireflection coatings, not shown in Fig. 2, is black layer coatings. Such coatings do not transmit or reflect any of the incident light in a given spectral region but absorb all the incident light. Such coatings find uses in high-contrast display devices and energy conversion coatings.

High-reflection coatings. These coatings are used whenever surfaces are required to have a high reflectance. At one time, metallic surfaces were the only way to achieve this goal; that is, metal substrates were polished until they formed mirrors. More recently, opaque metallic layers have been deposited onto polished glass surfaces to achieve the same effect. Some typical spectral reflectance curves of metals are shown in **Fig. 3**. They have broad regions of high reflectivity. However, when the highest reflectances are required, interference effects in stacks of nonabsorbing multilayers are used. In this way, reflectances as high as 99.9998% can be achieved in narrow parts of the visible and near-infrared spectral regions. Such values are required for some laser applications. Using special designs consisting of a larger number of layers of suitable thicknesses, high reflectances can also be achieved over broader spectral regions. For femtosecond lasers, reflectors are required whose phase dispersions must meet certain criteria. High-reflectance coatings are made for wavelengths ranging from the x-ray to the submillimeter spectral regions, although the reflectances that can be achieved depend greatly on the optical constants of the coating materials available in a given region. The performances of some typical high-reflectance

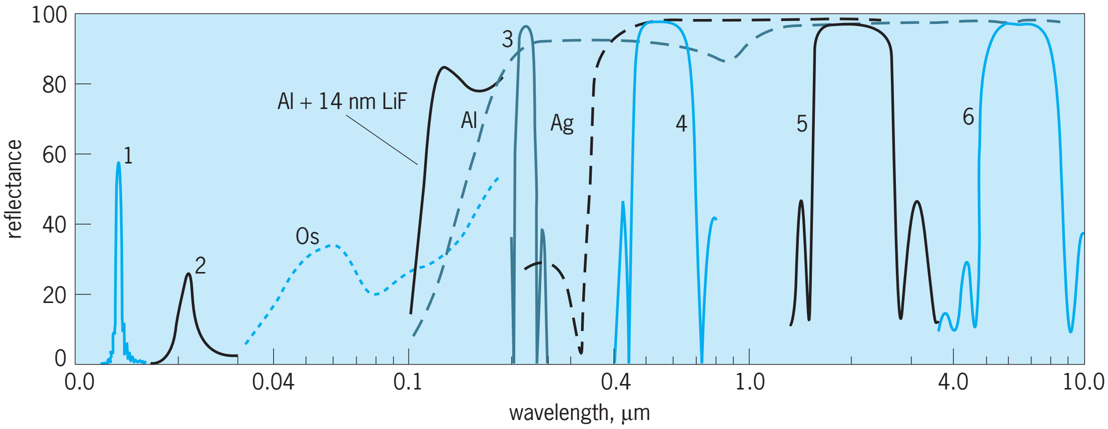


Fig. 3. Reflecting coatings. The solid curves represent typical measured reflectances of multilayer coatings for the x-ray (1, 2), ultraviolet (3), visible (4), and infrared (5, 6) spectral regions. The reflectances of opaque osmium, aluminum, and silver films (broken-line curves) are shown for comparison. The reflectances of metal layers can be extended or enhanced by additional thin-film coatings (for example, opaque aluminum with a 14-nanometer-thick LiF layer).

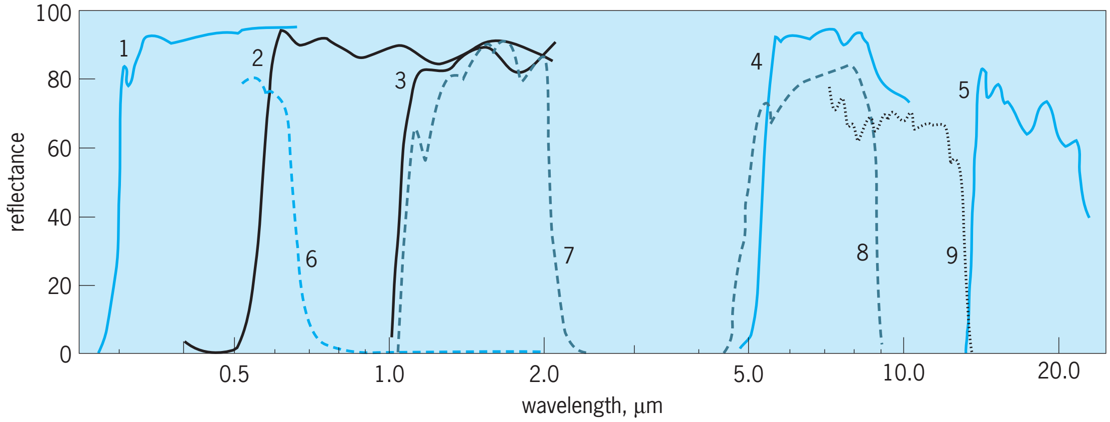


Fig. 4. Measured performances of short-wavelength (curves 1–5) and long-wavelength (6–9) cut-off filters for the ultraviolet, visible, and infrared spectral regions.

coatings for different wavelengths are shown in **Fig. 3**.

Long- and short-wavelength cut-off filters. These cut-off filters have an important technological application. For the visible and adjoining spectral regions, short-wavelength cut-off filters can be more easily produced from colored glasses. However, long-wavelength cut-off filters are difficult to produce except by thin-film interference. Even with this technique, the spectral ranges over which the transmittance remains high and the rejection low leave much to be desired. The performance of typical thin-film interference short- and long-wavelength cut-off filters is shown in **Fig. 4**.

Achromatic and color-selective beam splitters. Beam splitters are used in many scientific instruments, as well as in some consumer products. Suitable thin-film systems are deposited onto a plate or are located between two right-angled prisms. They split the obliquely incident light beam into two beams, one of which is transmitted and the other reflected. Usually when light is incident onto a multilayer system at nonnormal incidence, the transmittances and reflectances will be different for incident light that is polarized parallel (T_p, R_p) and perpendicular (T_s, R_s) to the plane of incidence. For beam splitters, this is a disadvantage and measures are often taken to reduce the effect. When the two beams have a uniform energy distribution in the spectral region of interest, the beam splitters are said to be achromatic. The beams may be of equal intensity, or one beam may be more intense than the other. Examples of the performance of such coatings are shown in **Fig. 5**. At other times, beam splitters are required in which the incident light is divided into beams of different color. Such beam splitters are called color-selective, and they are used in cameras and projection equipment for television and cinematography. The measured mean transmittances $(T_p + T_s)/2$ of one set of color selective beam splitters are shown in Fig. 5*c*.

Narrow-band-transmission interference filters. These are used in many scientific and engineering applications. In general, the specifications for narrow-band filters not only include the center wavelength and width at the points where transmittance is half of the maximum transmittance and the width of the rejection region where the reflectance is low, but also the widths at 0.1 and 0.01 of the maximum transmittance. The relative values of these widths are a measure of the

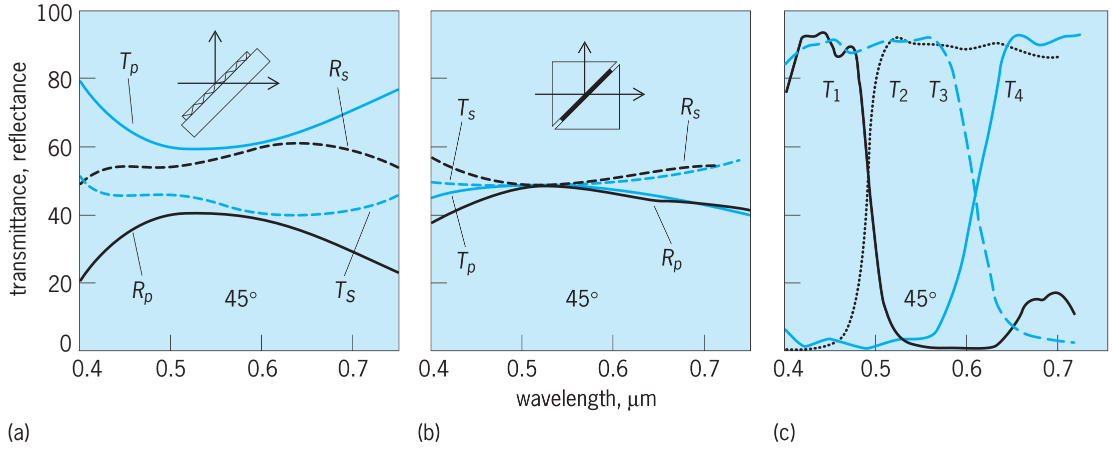


Fig. 5. Measured performance of (*a*) achromatic plate, (*b*) cube, (*c*) and color-selective beam splitters. In each case, the light was incident at 45° to the multilayer. The transmittance and reflectance may exhibit a strong dependence on the polarization of the incident light, as in *a* and *b*.

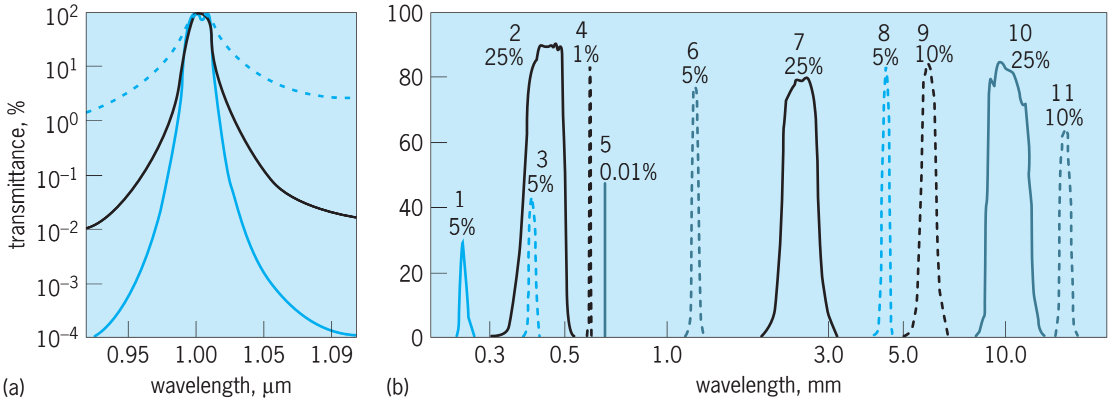


Fig. 6. Narrow-band transmission filters. (*a*) Calculated transmittance curves of filters having progressively squarer transmission bands. (*b*) Measured transmittances of a number of typical narrow-band interference filters for the ultraviolet, visible, and infrared spectral regions. The widths of the filters at half-maximum transmission, expressed as a percentage of the center wavelength, are indicated. All these filters do not transmit at shorter and longer wavelengths, except for the narrowest filter (curve 5), which requires additional blocking.

squareness of the transmission band (**Fig. 6*a***). The specifications are most stringent for filters for telecom applications, where the transmission in the rejection region may have to be as small as 10^{-5} and the transmission in the pass band may have to be in excess of 98%. To meet such stringent requirements, filters are required that often consist of more than 150 layers. A few representative narrow-band filters are shown in Fig. 6*b*. Narrow-band filters can be constructed with half-widths as narrow as 1 nanometer or as wide as 25% of the center wavelength.

Narrow-band rejection filters. Sometimes called minus filters, these filters should transmit all the incident radiation except for one or more specified wavelengths. Such filters are used in safety goggles for scientists and technologists working with lasers or for rejecting specific spectral lines in spectroscopic laboratories. Narrow-band rejection filters can consist of many layers made of two materials of close refractive indices, or they can consist of an inhomogeneous layer in which the refractive index oscillates gradually between two close values. Such coatings, often called rugate filters, have the advantage that the transmittance stays high over a broader range of wavelengths on the short-wavelength side of the rejection band. Examples of the performance of narrow-band rejection filters are shown in **Fig. 7**.

Polarizers and polarizing beam splitters. These find application in scientific instruments, as well as in liquid-crystal projection displays. Polarizers allow only light of one polarization to pass and block the light of the orthogonal polarization, while polarizing beam splitters separate unpolarized light into two orthogonally polarized transmitted and reflected beams of light. The purity of the polarized beam is measured by the extinction ratio, for example T_p/T_s or R_s/R_p, depending on whether the polarizer acts in transmission or reflection. Depending on the application, the polarizing devices might need to be effective over a narrow or broad spectral region. Most practical polarizers and polarizing beam splitters are based on isotropic interference multilayer coatings and need to operate with light incident at a specified oblique angle θ onto the multilayer. Most thin-film polarizers and polarizing beam splitters are effective only for light that is incident within a small angular field $\delta\theta_{\mathrm{air}}$ around θ, or operate only in a narrow spectral

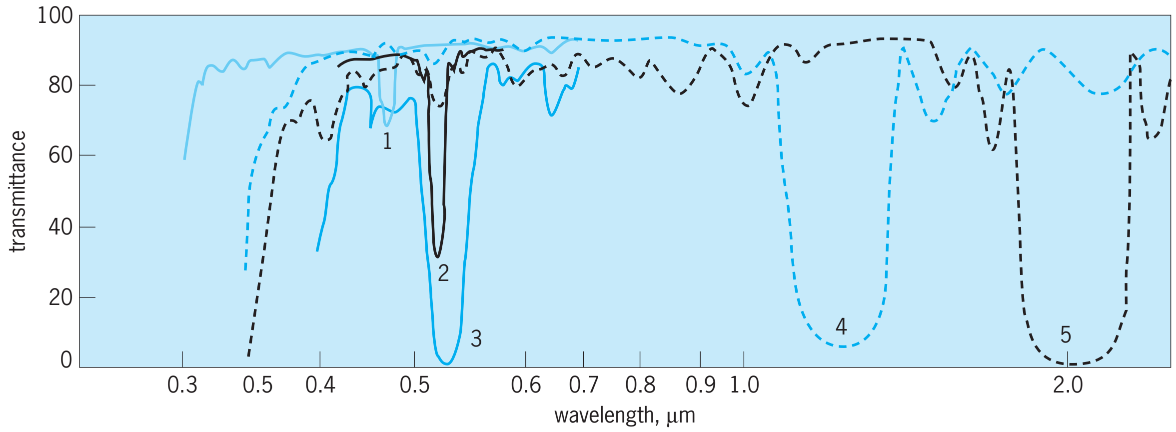


Fig. 7. Measured performance of narrow-band rejection filters based on homogeneous multilayers (curves 1–3) and inhomogeneous layers (4, 5).

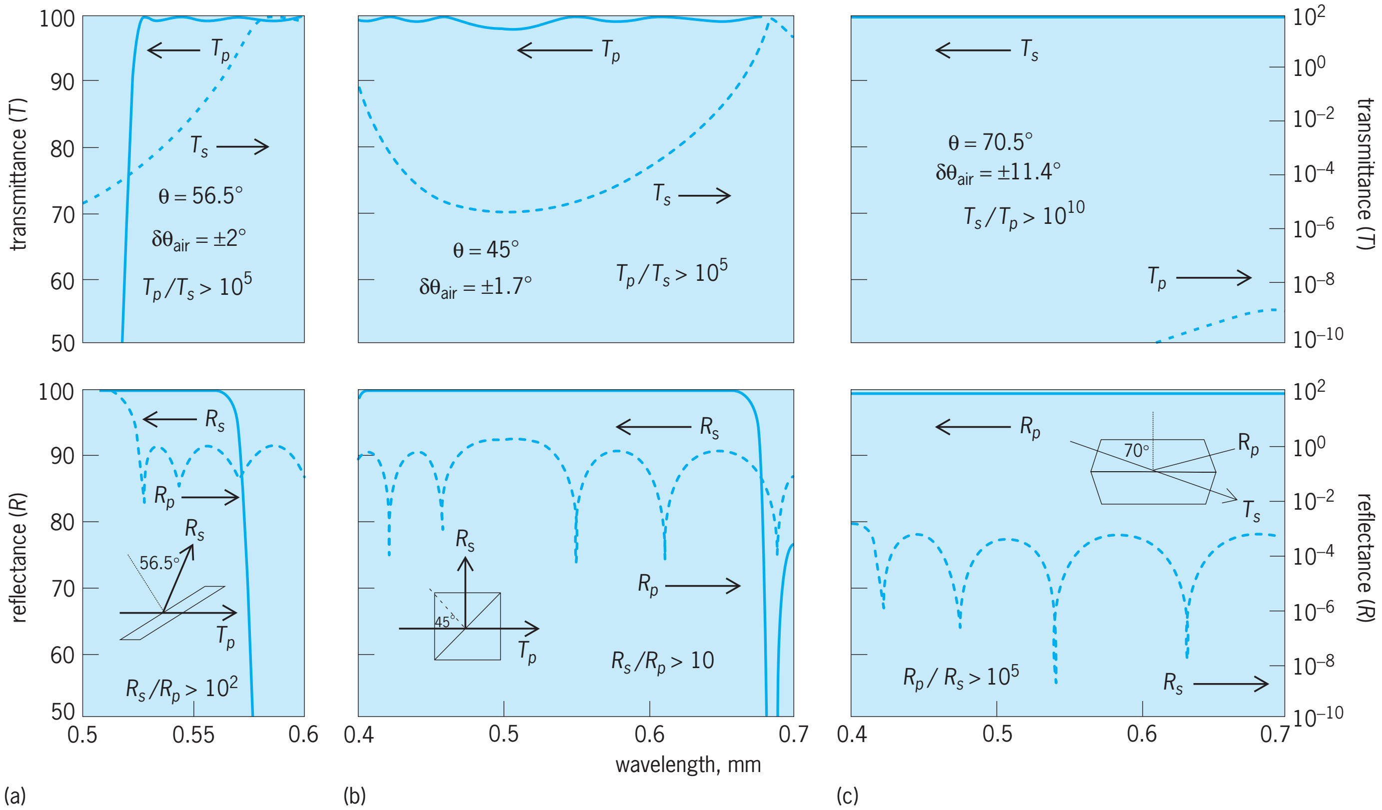


Fig. 8. Calculated performance of polarizers and polarizing beam splitters at the design angle θ. (*a*) Plate polarizers operate in a narrow spectral region with well-collimated incident light and can withstand high laser power. (*b*) MacNeille polarizing beam splitters operate over a broad spectral range, but their performances deteriorate rapidly with the departure of the incident beam from the design angle. (*c*) The Li-Li polarizing beam splitter operates over a broad spectral range and has a wide angular field for both the transmitted and reflected beams.

region. An exception is the Li-Li polarizing beam splitter, which is not yet (mid-2006) available commercially. The calculated spectral transmittances T_p, T_s and reflectances R_p, R_s of three types of polarizers and polarizing beam splitters at the design angles of incidence are shown in **Fig. 8**. Also given in the diagrams are the approximate extinction ratios for the transmitted and reflected beams at the edges of their angular fields measured in air. *See* POLARIZATION OF WAVES; POLARIZED LIGHT.

Correction and gain-flattening filters. Correction and gain-flattening filters are used when the spectral energy distribution of the light source or sensitivity of a detector vary too much for the satisfactory operation of a spectral or telecommunications device. The spectral transmittance of a correction or gain-flattening filter is such that in combination with the source and the detector, it will produce a more uniform response in the spectral region of interest. Correction filters often need to have very irregular spectral transmittances with values varying by orders of magnitude, especially when the light sources used have sharp emission lines.

State of the art. The thin-film interference filters and coatings whose properties are shown in Figs. 2–8 are only a few examples of what can be produced, since filters with properties intermediate to those in the graphs are possible. However, the performance that can be achieved depends also on the spectral region. At the outer edges of the spectrum for which interference filters are designed, the performance can be much worse than that in the visible and near infrared because of the lack of suitable coating materials.

The state of the art of thin-film optical interference filters has made great strides since the 1950s. It is now possible to design filters with almost any spectral transmittance or reflectance characteristics for normally incident light, provided that the spectral region over which they are specified is not too wide. Deposition processes exist now for the manufacture of very stable, mechanically durable coatings. Process controls have been developed that make it possible to produce quite complex multilayer coatings automatically. The accuracy of experimentally produced thin-film interference filters is limited by the accuracy of the measurements made during their deposition. In the future, the use of nanostructured films, with properties that supplement those found in natural materials, should lead to filters with improved properties. J. A. Dobrowolski

Bibliography. P. W. Baumeister, *Optical Coating Technology*, 2004; J. A. Dobrowolski, Optical properties of films and coatings, in *Handbook of Optics*, vol. 1, 2d ed., ed. by M. Bass, pp. 42.1–42.130, 1995; H. A. Macleod, *Thin-film Optical Filters*, 3d ed., 2001; J. D. Rancourt, *Optical Thin Films Users' Handbook*, 1987.

Interference-free circuitry

Circuitry that is designed to ensure that external sources of interference do not degrade its performance. The design principles involved are very simple, and their application in a systematic manner will reduce interference from external sources to an acceptable level, even for the most sensitive circuits and applications.

Electrical currents flow around complete, closed paths. These paths contain the energy sources which generate the currents. In complicated circuit networks comprising many meshes, a current from a given source may well divide and subdivide as it encounters circuit junction points, but these partial currents must recombine to give the original total current on return to a source. Unfortunately, most diagnostic equipment (such as oscilloscopes, voltmeters, and spectrum analyzers) responds to voltage differences, and the currents which cause these must be inferred. Nevertheless, attention must be concentrated on the currents flowing around circuits rather than voltages observed between points.

Coupling modes. Consider the simplest possible circuit containing an impedance and a voltage or current detector, as illustrated on the left of Figs. 1, 2, and 3. Suppose it is close to another circuit consisting of a source of interference and a load impedance, as on the right of these diagrams. There are only three ways in which interference from the source can couple in to the left-hand circuit and be registered as an unwanted signal by the output device D.

1. In **Fig. 1***a*, the two circuits have a portion of conductor A-B which is common to both. The current *I* in the right-hand circuit, which is the source of the interference, flows through this shared conductor, and this results in a voltage difference between A and B. This voltage in turn causes an unwanted interfering current to flow around the left-hand circuit, and to be registered by D. We could term this mode of coupling interference "common-conductor." It will be effective over a wide range of interfering frequencies and for both high and low values of the impedances in both circuits. Supposing the circuits must be connected at all, the measure to take to prevent interference coupling-in via this mode is to reduce the length of common conductor to a single point, as in Fig. 1*b*. If the circuits have to be spatially separated, a single conductor can be connected between the two circuits, as in Fig. 1*c*; and it is obviously still true that however large the current in the interfering circuit, no voltage difference or consequent interference current is produced in the left-hand circuit.

2. In **Fig. 2***a*, there is a connection between the circuits, and a stray capacitance *C* is shown between them. In Fig. 2*b*, there is no connection, but two stray capacitances are shown. Currents can circulate via capacitances to cause interference in the left-hand circuit. This capacitive coupling mode is usually significant at higher frequencies for circuits containing high impedances, say of the order of 100 Ω or more.

This coupling mode can be eliminated by strategically placed and connected conducting screens between the circuits, as shown in Fig. 2*c*.

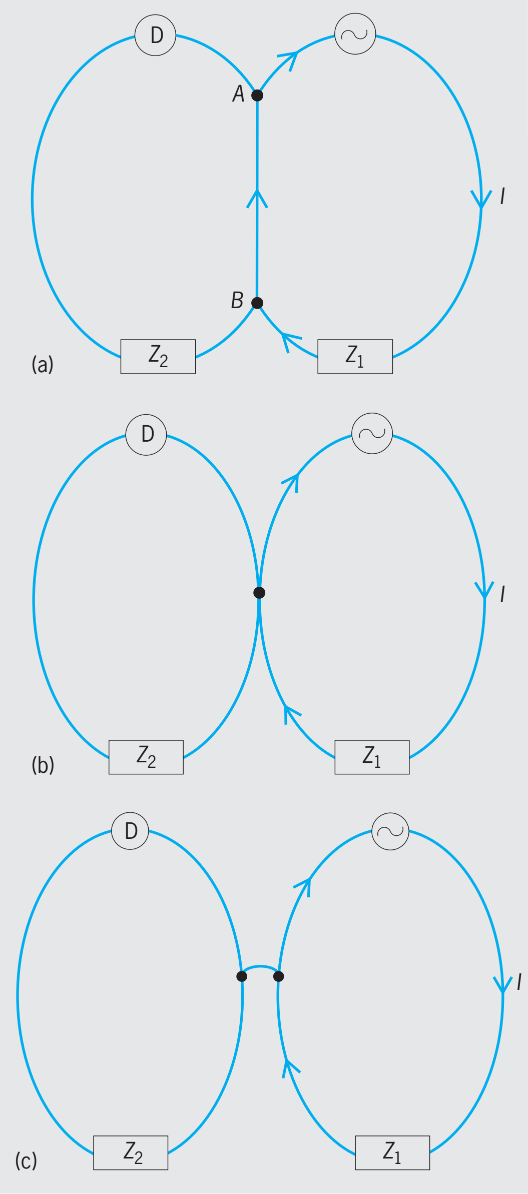


Fig. 1. Common-conductor coupling of two circuits. (*a*) Two circuits sharing a common conductor *AB*. *Z* represents load impedance, D is a detector, ∼ represents a source of interference, and *I* is current. Coupling is prevented by (*b*) reducing the common conductor to a point or (*c*) connecting the circuits with a length of conductor.

3. In **Fig. 3***a*, the circuits have no common connection, but some of the magnetic flux *B* created by the interfering circuit threads the left-hand circuit and induces an interfering emf in it and a consequent interference current around it. This mode of interference is most significant where the left-hand circuit is of low impedance, that is, of the order of 10,000 Ω or less.

The way to avoid this magnetic coupling mode is to reduce the coupling between the circuits by either increasing the distance between them or decreasing to a very small value the area enclosed by one or, preferably, both circuits. Figure 3*b*, *c*, and *d* illustrates three ways of accomplishing the latter. In Fig. 3*b*, the conductors of both circuits are routed close to each other in go-and-return pairs. In Fig. 3*c*, the area enclosed by the go-and-return pairs is further reduced by being twisted, with the advantage that the small flux generated is in opposing directions from successive loops of the twist and thereby cancels at a little distance from the circuit to an exceedingly small value. In Fig. 3*d*, the go-and-return conductors are arranged to be the inner and outer conductors of a coaxial cable. The line integral of magnetic flux around a concentric circular path of any radius outside the cable is zero because the total current threading it is zero. Because of the cross-sectional symmetry of a circular coaxial cable, this line integral can be zero only if the flux itself is zero everywhere. Although this result is strictly true only for a straight cable of infinite length, flux elimination is good enough for all practical purposes for short lengths of cable if the cable is not too sharply bent anywhere. A further advantage of the coaxial solution is that if, as is usual, the outer conductors of the sending and receiving circuits are at more or less the same potential and involve no appreciable impedance, they act as screens, and capacitive

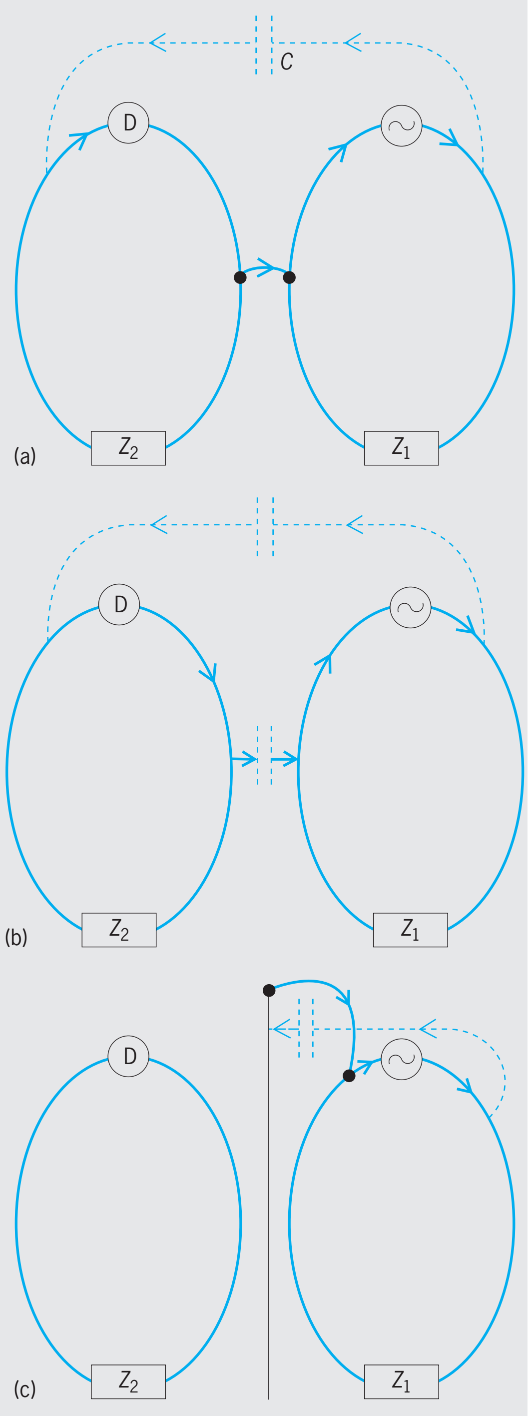


Fig. 2. Capacitive coupling of two circuits. (*a*) An interference current circulating via a stray capacitance *C*. (*b*) An interference current circulating via two stray capacitances. (*c*) Elimination of coupling by adding a screen to intercept capacitive currents.

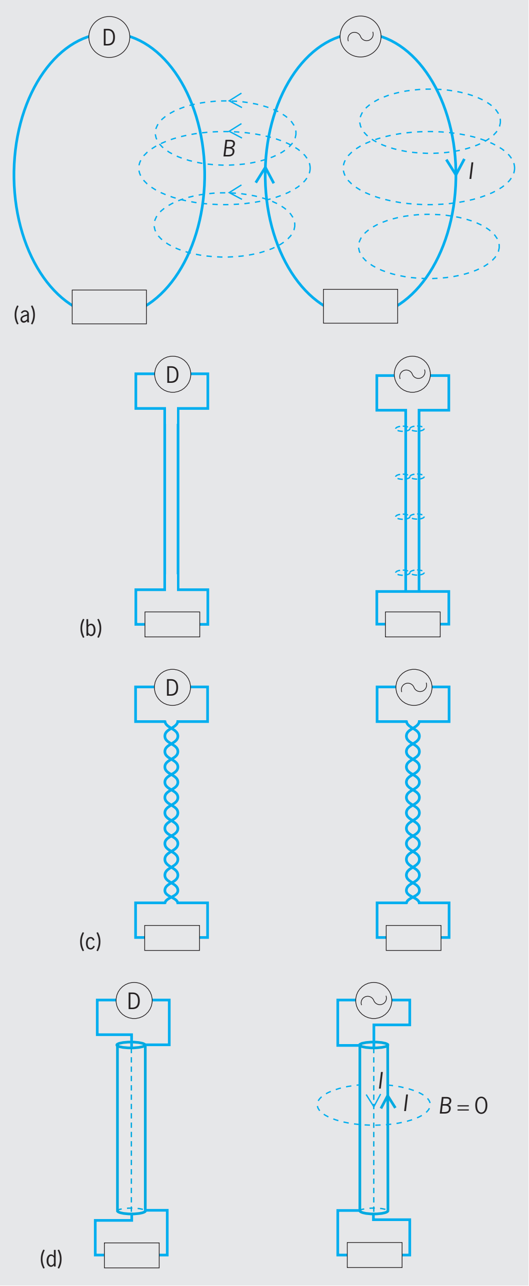


Fig. 3. Magnetic coupling of two circuits. (*a*) The magnetic field *B* created by the interference current *I* threads the left-hand circuit. (*b*) Decreasing the loop area of both circuits reduces magnetic coupling. (*c*) Twisting the conductors further reduces coupling. (*d*) An alternative solution employs coaxial connecting cables.

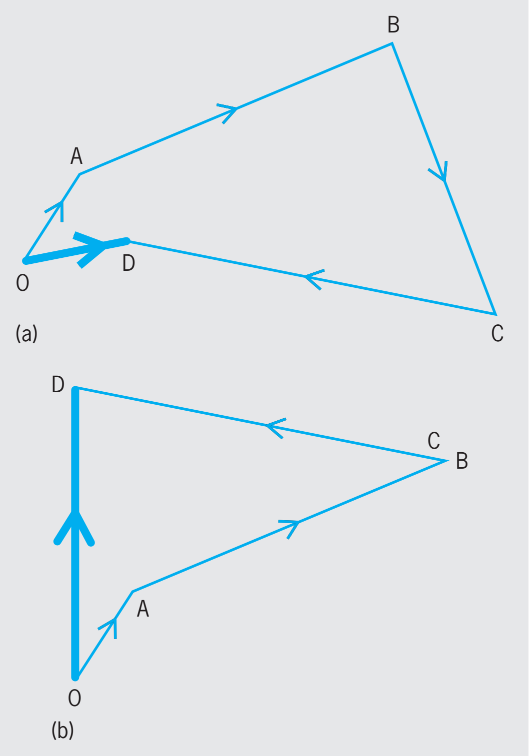


Fig. 4. Vector diagrams illustrating an unexpected result. (*a*) Summing vectors *OA*, *AB*, *BC*, and *CD* which represent interference at a single frequency provides the resultant *OD*. (*b*) The resultant *OD* is increased when the interference represented by *BC* is eliminated.

coupling is also reduced to negligible proportions. If, in an actual situation, the outer conductors are themselves interconnected as a network, at lower frequencies additional measures may need to be taken to ensure that the currents in the inner and outer conductors of each cable are equal and opposite. One way of doing this is to thread one cable in each mesh several times through a high-permeability magnetic core to ensure that the total impedance of the mesh is much greater than the go-and-return impedance. The latter is unaffected by the presence of the core. *See* COAXIAL CABLE; COUPLED CIRCUITS.

Need for systematic approach. In any actual situation, more than one of the three modes will be acting simultaneously, and in a complicated network there will be several coupling routes associated with each mode. Therefore a systematic approach is needed whereby each and every possible route for the ingress of interference is considered in turn and appropriate action taken.

Unfortunately, eliminating interference systematically in the common situation where interference at a single frequency enters circuitry through several routes or through more than one mode can cause puzzling results. An example of single-frequency coherent interference is that from mains or line frequency and its harmonics. In this case there may well be partial cancellation at the final output device of the interference entering via the various routes. This situation is illustrated in the vector diagram of **Fig.** 4*a*, where the vector representing the sum of the interference is *OD*. Then if one of the routes of entry is eliminated, say that corresponding to the vector *BC*, the remaining interference may well sum to larger value, as in Fig. 4*b*. This will make a trial-and-error approach impossible, and will force a more considered and systematic approach to be taken. Sometimes, coherent interference of this kind may well accidentally sum to an acceptably small value, or another route of adjustable phase and magnitude may be deliberately created to oppose the total interference. This situation is not a good one, because any change in the circumstances or environment of the circuitry will alter the magnitude and phase of the various coupling modes and the interference will reappear. Besides, interference in one part of sensitive circuitry may cause it to become saturated and to malfunction. Therefore the aim should be reduction to acceptable levels of interference via all possible routes. In the example of mains/line interference, this will have the added advantage that any other interference carried and propagated via mains/line circuitry will also be eliminated.

Isolation. Isolation is often an important concept in eliminating interference. An instrument or section of circuitry is said to be isolated if there are no routes by which significant currents, interfering or otherwise, can flow through the instrument or circuitry from a source onward to other circuitry. Circuitry entirely contained within a conduction enclosure apart from go-and-return current paths to supply power to it will be isolated. This description applies to many test and measurement instruments, even when connected to circuitry which is itself isolated, such as a resistance-measuring instrument connected to a resistance thermometer whose leads and sensing element are enclosed in a continuous shield. This shield should be connected only to the shield of the instrument, thus forming an extension of it. Unfortunately, isolation is usually compromised by the mains/line and power supply connections if the instrument or circuitry is connected into a larger, overall system of other instruments and circuits. One way of maintaining isolation in these circumstances is to make use of differential voltage-sensing input circuits which exhibit high input impedance. Battery-operated instruments whose batteries are contained entirely within the conducting case of the instrument can also offer a way out of some difficult situations. *See* ELECTRICAL SHIELDING.

Misconceptions. Finally, clear thinking about interference elimination will be aided if two common misconceptions are recognized and avoided.

1. Interference does not enter circuitry at localized points or circuit nodes. A complete current path must be involved.

2. For brevity in circuit descriptions and diagrams, many connections are said to be made to a conductor called ground. The term is best avoided, but if old habits die hard, at least it should be recognized that this name is shorthand for a whole network of conductors which return currents to their sources, thereby completing circuit loops for interference coupling modes. This network may well not be an equipotential conductor system because of the currents flowing in it, and may itself give

rise to common-conductor coupling, which can be avoided by the star or single-conductor connections described above. *See* ELECTRICAL INTERFERENCE; GROUNDING; ELECTROMAGNETIC COMPATIBILITY.

Bryan P. Kibble

Bibliography. *A Guide To Measuring Direct and Alternating Current and Voltage Below 1 MHz*, Institute of Measurement and Control, U.K., 2003; P. Horowitz and W. Hill, *The Art of Electronics*, 2d ed., Cambridge University Press, 1989; R. Morrison, *Grounding and Shielding Techniques in Instrumentation*, 3d ed., Wiley-Interscience, 1986.

Interference microscope

An instrument for visualizing and measuring differences in the phase of light transmitted through or reflected from microscopic specimens. It is closely allied to the phase-contrast microscope. *See* PHASE-CONTRAST MICROSCOPE.

Simple theory. A microscopic image is formed by the interaction of all light waves passing through the optical system. In a phase-contrast microscope, light diffracted by a transparent object is spatially separated from nondiffracted light in the back focal plane of the objective, where a phase plate alters the relative phases of the diffracted and nondiffracted beams so that they interfere in the image plane to produce a visible, intensity-modulated image. In an interference microscope, the diffracted and nondiffracted waves are not spatially separated, but light (the object beam) that passes through or is reflected from the object interferes with light (the reference beam) that passes through or is reflected from a different region of the specimen plane or is reflected from a comparison (reference) surface. For interference to be visible, the light beams must be coherent; in other words, the beams must maintain a constant relationship of wavelength, phase, and polarization over a relatively long period. The easiest way to achieve coherence is by using a device such as a semireflecting mirror, which splits a light beam into two beams. Random changes in the properties of successive photons from a given point in the source then affect both beams simultaneously. For a Mach-Zehnder type of interferometer (**Fig. 1**), light from a source is split into two beams, which after total reflection at the mirrors are recombined. If the optical paths are identical, the light waves arrive in phase at the point where an image is formed, and interfere constructively. A retarding specimen in one beam throws the beams out of phase and produces partial or complete destructive interference where the image is formed. Differences in optical path introduced by various parts of the object can be seen as variations in intensity or color. It may be noted that energy is conserved: light "missing" from the image appears at a right angle to it. The optical path difference introduced by the specimen expresses a relationship between the geometric thickness, and the refractive indices of the reference space and the specimen. The optical path difference can be measured by a calibrated compensator. It restores the relative phase of the two beams. *See* INTERFERENCE OF WAVES.

Many forms of interferometer have been described, but few have been used in commercial microscopes. A Leitz interference microscope using the Mach-Zehnder principle had two condensers, two objectives, and wide separation of the specimen and reference beams. Maintenance of the necessary mechanical rigidity was difficult and expensive, however, and so several compact, single-objective polarizing microinterferometers have been devised based on the Jamin-Lebedeff principle. In this type of instrument, plane-polarized light from a substage polarizer is split by a birefringent plate cemented to the condenser into two parallel beams polarized at right angles to each other and separated, or sheared, by a distance of about 30–200 micrometers, depending on the magnification of the objective used. After passing through the specimen plane, the beams are recombined by a birefringent plate cemented to the objective. The relative phase of the beams can be adjusted with a polarizing compensator such as a quartz wedge or a compensator that combines a fixed quarter-wave plate and a rotating analyzer. *See* BIREFRINGENCE.

Because of the beam splitters, an interference

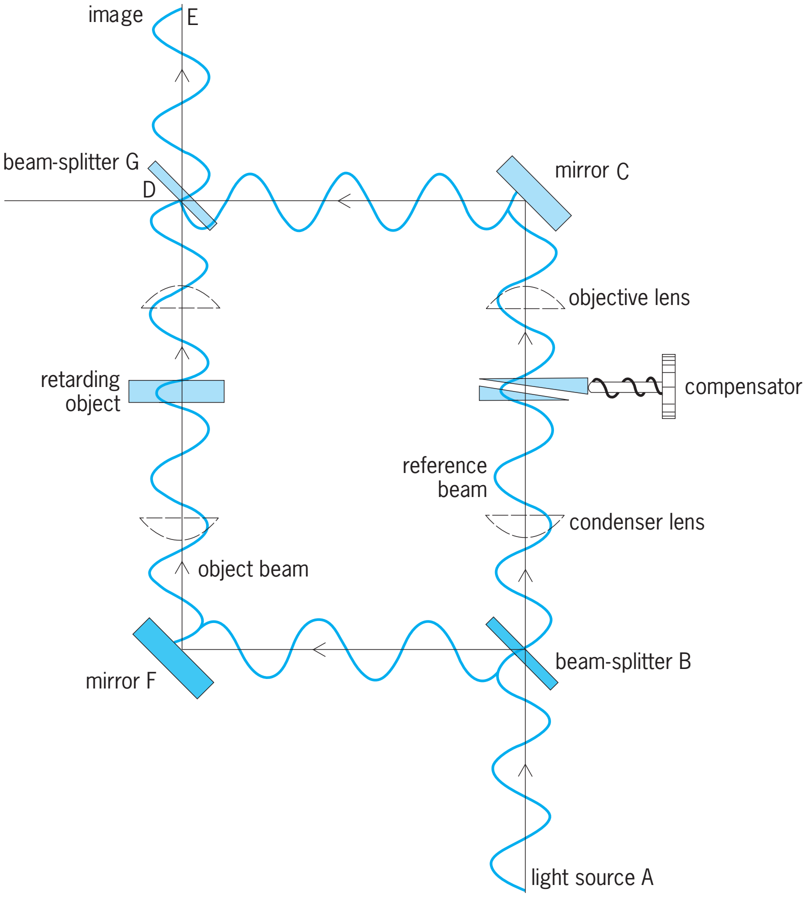


Fig. 1. Diagram of a Mach-Zehnder type of interference microscope (eyepiece not shown). Light from source A is split at B into coherent object and reference beams, which are reflected at F and C, respectively, and recombined at D. If the optical paths are identical, constructive interference occurs at E and complete destructive interference occurs at G. That relationship, if disturbed by a retarding specimen in the object beam, can be restored by a compensator in the reference beam. (*After D. J. Goldstein, Scanning microinterferometry, in G. A. Meek and H. Y. Elder, eds., Analytical and Quantitative Methods in Microscopy, Cambridge University Press, 1977*)

microscope displays two separate images of the specimen. In some instruments, one of these, the so-called ghost image, is astigmatic and out of focus. If the ordinary and ghost images overlap, as can occur with closely packed or large specimens, reliable measurements or observations in the overlapping region are impossible.

Measurement of phase changes. To measure the retardation of a relatively thick specimen, a gradient of retardation can be superimposed on the microscopic field by tilting an optical element or by inserting a retarding wedge in one beam. This produces a series of dark fringes across the field, uniformly spaced at intervals of one wavelength retardation. The deviation of fringes crossing a specimen is proportional to its retardation; whole wavelengths are readily counted, and fractions of a wavelength can be estimated with a precision of about 0.1 wavelength. To measure the retardation more precisely, a calibrated compensator (Fig. 1) is adjusted so that the background is maximally dark by using monochromatic light or some readily identifiable color such as sensitive purple. The compensator setting is recorded and then readjusted until the specimen appears with the same intensity or color as the background had. The difference between the two settings indicates the specimen retardation. The precision of visual setting can be improved appreciably by a half-shade device in the eyepiece, and under ideal circumstances is about 0.01 wavelength. With automatic microinterferometers that use mechanically phase-modulated light and a photoelectric detector, the reproducibility of repeated measurements of optical path difference exceeds 0.001 wavelength. The methods so far described measure the retardation at an arbitrary point in the object plane. In cell biology, however, it is often more useful to measure the integrated optical path difference of a discrete, nonhomogeneous specimen, which is the product of the specimen area and the mean optical path difference. The integrated optical path difference can be estimated by measuring the specimen area and by taking many measurements of local optical path differences or—more rapidly and precisely—by using a microinterferometer in its scanning and integrating mode.

Differential interference–contrast microscope. In Mach-Zehnder and Jamin-Lebedeff interference microscopes, the ordinary and ghost images of a discrete object are relatively widely separated, or sheared, and the specimen is compared with a clear region of background. In differential interference–contrast (DIC) instruments, on the other hand, the shear is very small and the images overlap almost completely; a given point of the specimen is compared not with empty mounting medium but with an adjacent object point, and contrast depends on the gradient of optical path in the direction of shear (**Fig. 2**). Consider, for example, a uniform sphere mounted in a medium of different refractive index. In the center of the image the optical path gradient is zero and the intensity is identical with the background. At one edge the gradient is maximal and the image is brighter than the background, whereas at the other edge the gradient is opposite in sign and the image is darker than the background. The object appears almost as if shadow-cast. With a given objective and a large condenser aperture, differential interference–contrast gives good lateral resolution and a small depth of focus; as a result, a degree of optical sectioning through the specimen can be achieved. Reducing the condenser aperture increases the contrast somewhat, but differential interference–contrast is less sensitive in general than phase contrast for detecting small phase differences. Note that in transmitted-light differential interference contrast, the contrast is related not to the gradient in geometrical object thickness but to the gradient in optical path, and so the image does not necessarily indicate true surface structure. For example, an internal cell organelle such as a nucleus, with a refractive index higher than that of the cytoplasm, may incorrectly seem to project from or to be an indentation in the surface of the cell. In spite of those interpretation problems, the striking three-dimensional appearance of differential interference contrast images has made the method popular in the field of biology.

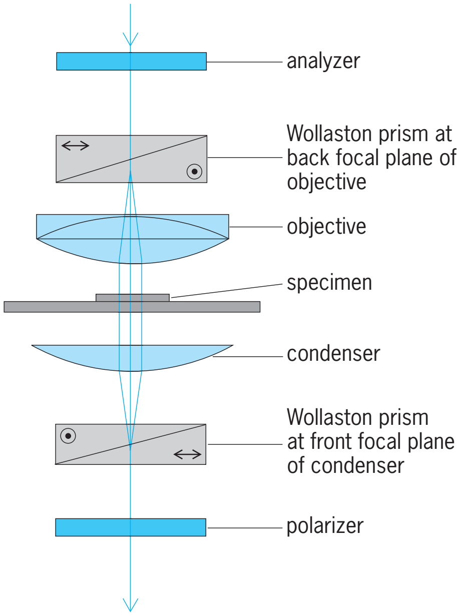


Fig. 2. Principle of a differential interference microscope with Wollaston prisms at the focal planes of the condenser and objective. Nomarski devised modified Wollaston prisms that can be used if the focal planes are not physically accessible. (*After W. Lang, Nomarski differential interference–contrast microscopy, Zeiss Inform., no. 70, pp. 114–120, 1969*)

Reflecting specimens. Surface structure can be studied with the Linnik microscope. If its mirror is tilted slightly relative to the reflecting surface, interference fringes appear and irregularities in the reflecting surface are seen as displacements in the fringe system. If the two surfaces are exactly parallel, the fringe will be of infinite width so that irregularities in the reflecting surface appear as intensity variations against a uniformly illuminated background, as in phase contrast. Nomarski differential interference contrast is also eminently suitable for reflected-light work. Excellent results have been obtained relatively inexpensively by using very narrow interference fringes produced by multiple reflection with an ordinary incident illuminator and a partially reflecting cover glass over the specimen.

Nomarski differential interference contrast is eminently suitable for reflected-light work. Also, there are phase-stepping or phase-shifting interference microscopes with which highly reproducible and accurate measurements of a surface can be made automatically and rapidly. This type of instrument typically uses a Mirau interference objective to illuminate the specimen with incident light, and a charge-coupled-device camera to capture an image with superimposed interference fringes. Intensities in the image are measured automatically, and the data are transferred to a computer. A small phase difference is introduced between the object and reference beams by moving either the reference surface or the specimen with a piezoelectric transducer controlled by the computer; this alters the fringes, and the intensities are again recorded. Provided at least three sets of intensity measurement are available, it is possible to compute the phase change at, and hence the height of, each image point. Some confocal scanning microscopes are able to function as microinterferometers, with the characteristic benefits of improved lateral resolution and optical sectioning ability given by confocal scanning. *See* CONFOCAL MICROSCOPY.

Comparison with phase-contrast microscope. The phase-contrast microscope can be regarded as an imperfect interference microscope in which the necessarily finite width of the phase annulus produces incomplete separation of the two interfering beams. The resulting halo effect and accentuation of edges are absent in some types of interference microscope. Although eliminating optical artifacts reduces the risk of incorrect image interpretation, it is not necessarily advantageous for purely observational work. Ghost images may hamper considerably the use of the interference microscope to view extended or densely packed specimens. In addition, the relative sensitivity, low cost, and ease of adjustment of the phase-contrast microscope make it the instrument of choice for routine observations and for measurement of refractive indices by the immersion method. The interference microscope is, however, essential for measuring phase change.

Biological applications. Knowing the thickness of the section being examined is often valuable in stereological work. Plastic sections more than 1 μm in thickness can readily be measured with a visual microinterferometer, and with an automatic instrument even ultrathin sections (about 50–70 nanometers) for electron microscopy can be measured with a reproducibility better than 5%.

The primary biological use of microinterferometry, however, is probably the determination of cellular dry mass. With an automatic scanning and integrating microinterferometer, the dry mass of a living mammalian cell can be precisely and nondestructively determined in a few seconds. Because measured optical path difference depends on the product of thickness and refractive index difference, only one of those quantities need be known to calculate the other. Thus the combination of interference microscopy and immersion microrefractometry is capable of yielding much useful quantitative information. In some cases it has proved possible to determine cellular volume, thickness, concentration of solids and water, and total dry and wet mass. Interference microscopy has also been successfully used in quantitative histochemistry for studying changes in dry mass produced by enzymic digestion or specific extraction procedures.

Nonbiological procedures. Like the phase-contrast microscope, the transmission interference microscope can be applied to the qualitative study of transparent specimens such as fibers and crystals. Incident-light instruments are employed in metallurgy, engineering, and the electronic industry to study specimens such as metallic surfaces, metalized replicas of surfaces, and electronic circuits. In that type of work it is a common practice to operate the instrument with fringes in the field of view, the shape of the fringes giving a contour map of the surface. The sensitivity is such that irregularities of the order of 0.01 wavelength or even less can be observed. The method gives valuable information about the quality of ground, polished, or etched surfaces. The differential (Nomarski) type of interference microscope has proved particularly useful for studying the surface structure of reflecting objects, since with that type of specimen it gives a true relief image uncomplicated by variations in refractive index. Vertical resolution on the order of 0.1 nm has been obtained with incident-light differential interference contrast microscopes combined with electronic image-analysis devices. *See* DIFFRACTION; OPTICAL MICROSCOPE; POLARIZED LIGHT; POLARIZED LIGHT MICROSCOPE; RESOLVING POWER (OPTICS).

D. J. Goldstein

Bibliography. M. Francon and S. Mallick, *Polarization Interferometers: Applications in Microscopy and Macroscopy*, 1971; W. Lang, Nomarski differential interference–contrast microscopy, *Zeiss Inform.*, nos. 70 and 71, 1969; G. A. Meek and H. Y. Elder (eds.), *Analytical and Quantitative Methods in Microscopy*, 1977; M. Pluta, *Advanced Light Microscopy*, vol. 3, 1993; S. Tolansky, *Multiple-Beam Interference Microscopy of Metals*, 1970.

Interference of waves

The process whereby two or more waves of the same frequency or wavelength combine to form a wave whose amplitude is the sum of the amplitudes of the interfering waves. The interfering waves can be electromagnetic, acoustic, or water waves, or in fact any periodic disturbance.

The most striking feature of interference is the effect of adding two waves in which the trough of one wave coincides with the peak of another. If the two waves are of equal amplitude, they can cancel each other out so that the resulting amplitude is zero. This is perhaps most dramatic in sound waves; it is possible to generate acoustic waves to arrive at a person's ear so as to cancel out disturbing noise. In optics, this cancellation can occur for particular wavelengths in a situation where white light is a source. The resulting light will appear colored. This gives rise to the iridescent colors of beetles' wings and mother-of-pearl, where the substances involved are actually colorless or transparent.

Two-beam interference. The quantitative features of the phenomenon can be demonstrated most easily by considering two interfering waves. The amplitude of the first wave at a particular point in space can be written as Eq. (1), where A_0 is the peak amplitude, and ω is 2π times the frequency. For the second wave Eq.(2) holds, where $\varphi_1 - \varphi_2$ is the phase difference between the two waves. In interference, the two waves are superimposed, and the resulting wave can be written as Eq. (3). Equation (3) can be expanded to give Eq. (4).

$$A = A_0 \sin(\omega t + \varphi_1) \tag{1}$$

$$B = B_0 \sin(\omega t + \varphi_2) \tag{2}$$

$$A + B = A_0 \sin(\omega t + \varphi_1) + B_0 \sin(\omega t + \varphi_2) \tag{3}$$

$$\begin{aligned} A + B = {} & (A_0 \sin\varphi_1 + B_0 \sin\varphi_2)\cos\omega t \\ & + (A_0 \cos\varphi_1 + B_0 \cos\varphi_2)\sin\omega t \end{aligned} \tag{4}$$

By writing Eqs. (5) and (6), Eq. (4) becomes Eq. (7), where C^2 is defined in Eq. (8). When C is less than A or B, the interference is called destructive. When it is greater, it is called constructive. For electromagnetic radiation, such as light, the amplitude in Eq. (7) represents an electric field strength. This field is a vector quantity and is associated with a particular direction in space, the direction being generally at right angles to the direction in which the wave is moving. These electric vectors can be added even when they are not parallel. For a discussion of the resulting interference phenomena *see* POLARIZED LIGHT; SUPERPOSITION PRINCIPLE.

$$A_0 \sin\varphi_1 + B_0 \sin\varphi_2 = C \sin\varphi_3 \tag{5}$$

$$A_0 \cos\varphi_1 + B_0 \cos\varphi_2 = C \cos\varphi_3 \tag{6}$$

$$A + B = C \sin(\omega t + \varphi_3) \tag{7}$$

$$C^2 = {A_0}^2 + {B_0}^2 + 2A_0B_0 \cos(\varphi_2 - \varphi_1) \tag{8}$$

In the case of radio waves or microwaves which are generated with vacuum tube or solid-state oscillators, the frequency requirement for interference is easily met. In the case of light waves, it is more difficult. Here the sources are generally radiating atoms. The smallest frequency spread from such a light source will still have a bandwidth of the order of 10^7 Hz. Such a bandwidth occurs in a single spectrum line, and can be considered a result of the existence of wave trains no longer than 10^{-8} s. The frequency spread associated with such a pulse can be written as notation (9), where t is the pulse length. This means

$$\Delta f \simeq \frac{1}{2\pi t} \tag{9}$$

that the amplitude and phase of the wave which is the sum of the waves from two such sources will shift at random in times shorter than 10^{-8} s. In addition, the direction of the electric vector will shift in these same time intervals. Light which has such a random direction for the electric vector is termed unpolarized. When the phase shifts and direction changes of the light vectors from two sources are identical, the sources are termed coherent.

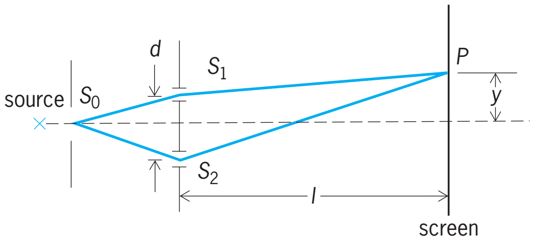


Fig. 1. Young's two-slit interference.

Splitting of light sources. To observe interference with waves generated by atomic or molecular transitions, it is necessary to use a single source and to split the light from the source into parts which can then be recombined. In this case, the amplitude and phase changes occur simultaneously in each of the parts at the same time.

Young's two-slit experiment. The simplest technique for producting a splitting from a single source was done by T. Young in 1801 and was one of the first demonstrations of the wave nature of light. In this experiment, a narrow slit is illuminated by a source, and the light from this slit is caused to illuminate two adjacent slits. The light from these two parallel slits can interfere, and the interference can be seen by letting the light from the two slits fall on a white screen. The screen will be covered with a series of parallel fringes. The location of these fringes can be derived approximately as follows: In **Fig. 1**, S_1 and S_2 are the two slits separated by a distance d. Their plane is a distance l from the screen. Since the slit S_0 is equidistant from S_1 and S_2, the intensity and phase of the light at each slit will be the same. The light falling on P from slit S_1 can be represented by Eq. (10) and from S_2 by Eq. (11), where f is the frequency, t the time, c the velocity of light; x_1 and x_2 are the distances of P from S_1 and S_2, and A_0 is the amplitude. This amplitude is assumed to be the same for each wave since the slits are close together, and x_1 and x_2 are thus nearly the same. These equations are the same as Eqs. (1) and (2), with $\varphi_1 = x_1/c$ and $\varphi_2 = x_2/c$. Accordingly, the square of the amplitude or the

$$A = A_0 \sin 2\pi f \left(t - \frac{x_1}{c}\right) \tag{10}$$

$$B = A_0 \sin 2\pi f \left(t - \frac{x_2}{c}\right) \tag{11}$$

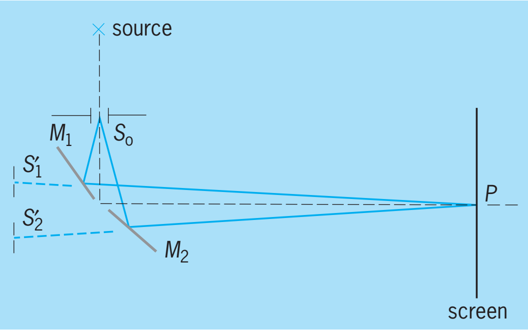


Fig. 2. Fresnel's double-mirror interference.

intensity at P can be written as Eq. (12).

$$I = 4A_0^2\cos^2\frac{2\pi f}{c}(x_1 - x_2) \qquad (12)$$

In general, l is very much larger than y so that Eq. (12) can be simplified to Eq. (13).

$$I = 4A_0^2\cos^2\pi\left(\frac{yd}{l\lambda}\right) \qquad (13)$$

Equation (13) is a maximum when Eq. (14) holds and a minimum when Eq. (15) holds, where n is an integer.

$$y = n\lambda\frac{l}{d} \qquad (14)$$

$$y = (n + {}^1/_2)\lambda\frac{l}{d} \qquad (15)$$

Accordingly, the screen is covered with a series of light and dark bands called interference fringes. If the source behind slit S_0 is white light and thus has wavelengths varying perhaps from 400 to 700 nm, the fringes are visible only where $x_1 - x_2$ is a few wavelengths, that is, where n is small. At large values of n, the position of the nth fringe for red light will be very different from the position for blue light, and the fringes will blend together and be washed out. With monochromatic light, the fringes will be visible out of values of n which are determined by the diffraction pattern of the slits. For an explanation of this *see* DIFFRACTION.

The energy carried by a wave is measured by the intensity, which is equal to the square of the amplitude. In the preceding example of the superposition of two waves, the intensity of the individual waves in Eqs. (1) and (2) is A^2 and B^2, respectively. When the phase shift between them is zero, the intensity of the resulting wave is given by Eq. (16).

$$(A + B)^2 = A^2 + 2AB + B^2 \qquad (16)$$

This would seem to be a violation of the first law of thermodynamics, since this is greater than the sum of the individual intensities. In any specific experiment, however, it turns out that the energy from the source is merely redistributed in space. The excess energy which appears where the interference is constructive will disappear in those places where the energy is destructive. This is illustrated by the fringe pattern in the Young two-slit experiment. The energy on the screen from each slit alone is given by Eq. (17), where A_0^2 is the intensity of the light from

$$E_1 = \int_0^\infty A_0^2\,dy \qquad (17)$$

each slit as given by Eq. (10). The intensity from the two slits without interference would be twice this value. The intensity with interference is given by Eq. (18). The comparison between $2E_1$ and E_3 need

$$E_3 = \int_0^\infty 4A^2\cos^2\left[2\pi\left(\frac{yd}{l\lambda}\right)\right]dy \qquad (18)$$

be made only over a range corresponding to one full cycle of fringes. This means that the argument of the cosine in Eq. (18) need be taken only from zero to π. This corresponds to a section of screen going from the center to a distance $y = l\lambda/2d$. From the two slits individually, the energy in this section of screen can be written as Eq. (19).

$$2E_1 = 2\int_0^{l\lambda/2d} A_0^2\,dy = \frac{A_0^2 l\lambda}{d} \qquad (19)$$

With interference, the energy is given by Eq. (20). Equation (20) can be written as Eq. (21).

$$E_3 = \int_0^{l\lambda/2d} 4A_0^2\cos^2\left[\left(\frac{yd}{l\lambda}\right)\right]dy \qquad (20)$$

$$E_3 = \frac{l\lambda}{2\pi d}\int_0^\pi 4A_0^2\cos^2\varphi\,d\varphi = \frac{A_0^2 l\lambda}{d} \qquad (21)$$

Thus, the total energy falling on the screen is not changed by the presence of interference. The energy density at a particular point is, however, drastically changed. This fact is most important for those waves of the electromagnetic spectrum which can be generated by vacuum-tube oscillators. The sources of radiation or antennas can be made to emit coherent waves which will undergo interference. This makes possible a redistribution of the radiated energy. Quite narrow beams of radiation can be produced by the proper phasing of a linear antenna array. *See* ANTENNA (ELECTROMAGNETISM).

The double-slit experiment also provides a good illustration of Niels Bohr's principle of complementarity. For detailed information on this *see* QUANTUM MECHANICS

Fresnel double mirror. Another way of splitting the light from the source is the Fresnel double mirror (**Fig. 2**). Light from the slit S_0 falls on two mirrors M_1 and M_2 which are inclined to each other at an angle of the order of a degree. On a screen where the illumination from the two mirrors overlaps, there will appear a set of interference fringes. These are the same as the fringes produced in the two-slit experiment, since the light on the screen comes from the images of the slits S_1' and S_2' formed by the two mirrors, and these two images are the equivalent of two slits.

Fresnel biprism. A third way of splitting the source is the Fresnel biprism. A sketch of a cross section of this device is shown in **Fig. 3**. The light from the slit at S_0 is transmitted through the two halves of the prism to the screen. The beam from each half will strike the

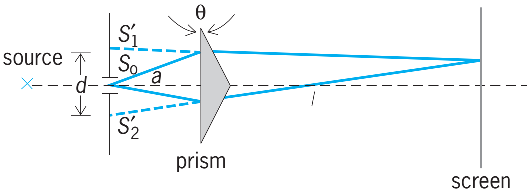


Fig. 3. Fresnel biprism interference.

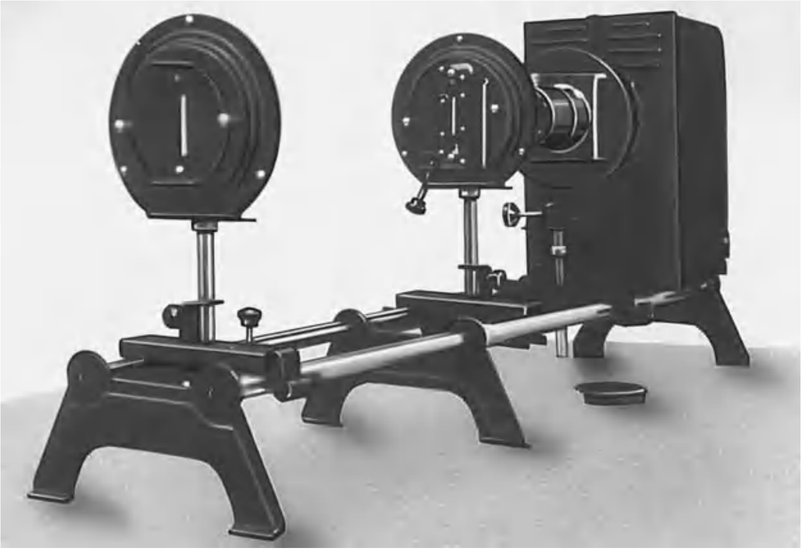


Fig. 4. Equipment for demonstrating Fresnel biprism interference.

screen at a different angle and will appear to come from a source which is slightly displaced from the original slit. These two virtual slits are shown in the sketch at S'_1 and S'_2. Their separation will depend on the distance of the prism from the slit S_0 and on the angle θ and index of refraction of the prism material. In Fig. 3, a is the distance of the slit from the biprism, and l the distance of the biprism from the screen. The distance of the two virtual slits from the screen is thus $a + l$. The separation of the two virtual slits is given by Eq. (22), where μ is the refractive index

$$d = 2a(\mu - 1)\theta \tag{22}$$

of the prism material. This can be put in Eq. (14) for the two-slit interference pattern to give Eq. (23) for the position of a bright fringe.

$$y = n\lambda \frac{a + l}{2a(\mu - 1)\theta} \tag{23}$$

A photograph of the experimental equipment for demonstrating interference with the Fresnel biprism is shown in **Fig. 4**. A typical fringe pattern is shown in **Fig. 5**. This pattern was obtained with a mercury-arc source, which has several strong spectrum lines, accounting in part for the intensity variation in the pattern. The pattern is also modified by diffraction at the apex of the prism.

Billet split lens. The source can also be split with the Billet split lens (**Fig. 6**). Here a simple lens is sawed into two parts which are slightly separated.

Lloyd's mirror. An important technique of splitting the source is with Lloyd's mirror (**Fig. 7**). The slit S_1 and its virtual image S'_2 constitute the double source. Part of the light falls directly on the screen, and part is reflected at grazing incidence from a plane mirror. This experiment differs from the previously discussed experiments in that the two beams are no longer identical. If the screen is moved to a point where it is nearly in contact with the mirror, the fringe of zero path difference will lie on the intersection of the mirror plane with the screen. This fringe turns out to be dark rather than light, as in the case of the previous interference experiments. The only explanation for this result is that light experiences a 180° phase shift on reflection from a material of higher refractive index than its surrounding medium. The equation for maximum and minimum light intensity at the screen must thus be interchanged for Lloyd's mirror fringes.

Amplitude splitting. The interference experiments discussed have all been done by splitting the wavefront of the light coming from the source. The energy from the source can also be split in amplitude. With such amplitude-splitting techniques, the light from the source falls on a surface which is partially reflecting. Part of the light is transmitted, part is reflected, and after further manipulation these parts are recombined to give the interference. In one type of experiment, the light transmitted through the surface is reflected from a second surface back through the partially reflecting surface, where it combines with the wave reflected from the first surface (**Fig. 8**). Here the arrows represent the normal to the wavefront of the light passing through surface S_1 to surface S_2. The wave is incident at A and C. The section

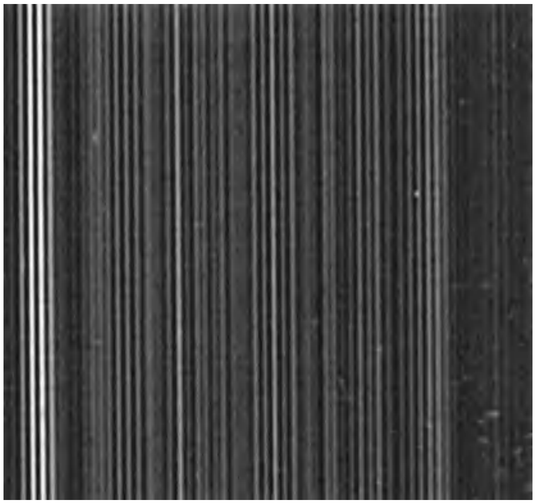


Fig. 5. Interference fringes formed with Fresnel biprism and mercury-arc light source.

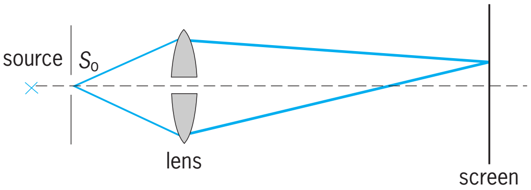


Fig. 6. Billet split-lens interference.

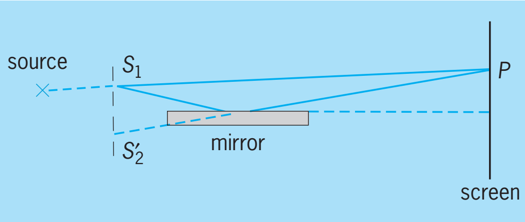


Fig. 7. Lloyd's mirror interference.

at A is partially transmitted to B, where it is again partially reflected to C. The wave leaving C now consists of two parts, one of which has traveled a longer distance than the other. These two waves will interfere. Let AD be the perpendicular from the ray at A to the ray going to C. The path difference will be given by Eq. (24), where μ is the refractive index of the

$$\Delta = 2\mu(AB) - (CD) \tag{24}$$

medium between the surfaces S_1 and S_2, and AB and CD are defined in Eqs. (25) and (26).

$$(AB) = \frac{d}{\cos\ r} \tag{25}$$

$$(CD) = 2(AB)\sin r \cos i \tag{26}$$

From Snell's law, Eq. (27) is obtained and thus Eq. (28) holds.

$$\sin i = \mu \sin r \tag{27}$$

$$\Delta = \frac{2\mu d}{\cos r} - \frac{2\mu d}{\cos r}\sin^2 r = 2\mu d \cos r \tag{28}$$

The difference in terms of wavelength and the phase difference are, respectively, given by Eqs. (29) and (30).

$$\Delta' = \frac{2\mu d \cos r}{\lambda} \tag{29}$$

$$\Delta\varphi = \frac{4\pi\mu d\ \cos r}{\lambda} + \pi \tag{30}$$

The phase difference of π radians is added because of the phase shift experienced by the light reflected at S_1. The experimental proof of this 180° phase shift was shown in the description of interference with Lloyd's mirror. If the plate of material has a lower index than the medium in which it is immersed, the π radians must still be added in Eq. (30), since now the beam reflected at S_2 will experience this extra phase shift. The purely pragmatic necessity of such an additional phase shift can be seen by considering the intensity of the reflected light when the surfaces S_1 and S_2 almost coincide. Without the extra phase shift, the two reflected beams would be in phase and the reflection would be strong. This is certainly not proper for a film of vanishing thickness. Constructive interference will take place at wavelengths for which $\Delta\varphi = 2m\pi$, where m is an integer. If the surfaces S_1 and S_2 are parallel, the fringes will be located optically at infinity. If they are not parallel, d will be a function of position along the surfaces and the fringes will be located near the surface. The intensity of the fringes will depend on the value of the partial reflectivity of the surfaces.

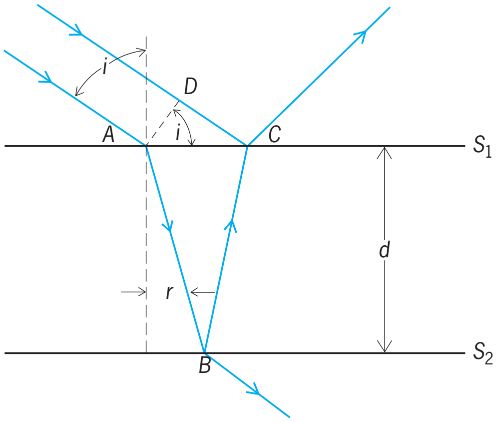


Fig. 8. Dielectric-plate-reflection interference.

Testing of optical surfaces. Observation of fringes of this type can be used to determine the contour of a surface. The surface to be tested is put close to an optically flat plate. Monochromatic light is reflected from the two surfaces and examined as in Fig. 8. One of the first experiments with fringes of this type was performed by Isaac Newton. A convex lens is pressed against a glass plate and illuminated with monochromatic light. A series of circular interference fringes known as Newton's rings appear around the point of contact. From the separation between the fringes, it is possible to determine the radius of curvature of the lens.

Thin films. Interference fringes of this two-surface type are responsible for the colors which appear in oil films floating on water. Here the two surfaces are the oil-air interface and the oil-water interface. The films are close to a visible light wavelength in thickness. If the thickness is such that, in a particular direction, destructive interference occurs for green light, red and blue will still be reflected and the film will have a strong purple appearance. This same general phenomenon is responsible for the colors of beetles' wings.

Channeled spectrum. Amplitude splitting shows clearly another condition that must be satisfied for interference to take place. The beams from the source must not only come from identical points, but they must also originate from these points at nearly the same time. The light which is reflected from C in Fig. 8 originates from the source later than the light which makes a double traversal between S_1 and S_2. If the surfaces are too far apart, the spectral regions of constructive and destructive interference become so close together that they cannot be resolved. In the case of interference by wavefront splitting, the light from different parts of a source could only be considered coherent if examined over a sufficiently short time interval. In the case of amplitude splitting, the interference when surfaces are widely separated can only be seen if examined over a sufficiently narrow frequency interval. If the two surfaces are illuminated with white light and the eye is used as the analyzer, interference cannot be seen when the separation is more than a few wavelengths. The interval between successive wavelengths of constructive interference becomes so small that each spectral region to which the eye is sensitive is illuminated, and no color is seen. In this case, the interference can again be seen by examining the reflected light with a spectroscope. The spectrum will be crossed with a set of dark fringes at those wavelengths for which there is destructive interference. This is called a channeled spectrum. For large separations of the surfaces, the separation between the wavelengths of destructive interference becomes smaller than the resolution of the spectrometer, and the fringes are no longer visible.

Fresnel coefficient. The amplitude of the light reflected at normal incidence from a dielectric surface is given by the Fresnel coefficient, Eq. (31), where

$$A = A_0 \frac{n_1 - n_2}{n_1 + n_2} \tag{31}$$

A_0 is the amplitude of the incident wave and n_1 and n_2 are the refractive indices of the materials in the order in which they are encountered by the light. In the simple case of a dielectric sheet, the intensity of the light reflected normally will be given by Eq. (32), where B is the amplitude of the wave which

$$C^2 = A^2 + B^2 + 2AB\ \cos\varphi \tag{32}$$

has passed through the sheet and is reflected from the second surface and back through the sheet to join A. The value of B is given by Eq. (33), where the

$$B = \frac{n_2 - n_3}{n_2 + n_3} \tag{33}$$

approximation is made that the intensity of the light is unchanged by passing through the first surface and where n_3 is the index of the material at the boundary of the far side of the sheet.

Nonreflecting film. An interesting application of Eq. (32) is the nonreflecting film. A single dielectric layer is evaporated onto a glass surface to reduce the reflectivity of the surface to the smallest possible value. From Eq. (32) it is clear that this takes place when $\cos\varphi = -1$. If the surface is used in an instrument with a broad spectral range, such as a visual device, the film thickness should be adjusted to put the interference minimum in the first order and in the middle of the desired spectral range. For the eye, this wavelength is approximately in the yellow so that such films reflect in the red and blue and appear purple. The index of the film should be chosen to make $C^2 = 0$. At this point Eqs. (34)–(36) hold.

$$(A - B)^2 = 0 \tag{34}$$

$$\frac{n_1 - n_2}{n_1 + n_2} = \frac{n_2 - n_3}{n_2 + n_3} \tag{35}$$

$$n_1 n_2 - n_2{}^2 + n_1 n_3 - n_2 n_3 = n_1 n_2 - n_1 n_3 + n_2{}^2 - n_2 n_3 \tag{36}$$

Equation (36) can be reduced to Eq. (37). In the

$$n_2 = \sqrt{n_1 n_3} \tag{37}$$

case of a glass surface in air, $n_1 = 1$ and $n_3 \cong 1.5$. Magnesium fluoride is a substance which is frequently used as a nonreflective coating, since it is hard and approximately satisfies the relationship of Eq. (37). The purpose of reducing the reflection from an optical element is to increase its transmission, since the energy which is not reflected is transmitted. In the case of a single element, this increase is not particularly important. Some optical instruments may have 15–20 air-glass surfaces, however, and the coating of these surfaces gives a tremendous increase in transmission.

Haidinger fringes. When the second surface in two-surface interference is partially reflecting, interference can also be observed in the wave transmitted through both surfaces. The interference fringes will be complementary to those appearing in reflection. Their location will depend on the parallelism of the surfaces. For plane parallel surfaces, the fringes will appear at infinity and will be concentric rings. These were first observed by W. K. Haidinger and are called Haidinger fringes.

Multiple-beam interference. If the surfaces S_1 and S_2 are strongly reflecting, it is necessary to consider multiple reflections between them. For air-glass surfaces, this does not apply since the reflectivity is of the order of 4%, and the twice-reflected beam is much reduced in intensity.

In **Fig. 9** the situation in which the surfaces S_1 and S_2 have reflectivities r_1 and r_2 is shown. The space between the surfaces has an index n_2 and thickness d. An incident light beam of amplitude A is partially reflected at the first surface. The transmitted component is reflected at S_2 and is reflected back to S_1 where a second splitting takes place. This is repeated. Each successive component of the waves leaving S_1 is retarded with respect to the next. The amount of each retardation is given by Eq. (38).

$$\varphi = \frac{4\pi nd}{\lambda} \cos\theta \tag{38}$$

Equation (7) was derived for the superposition of two waves. It is possible to derive a similar expression for the superposition of many waves. From

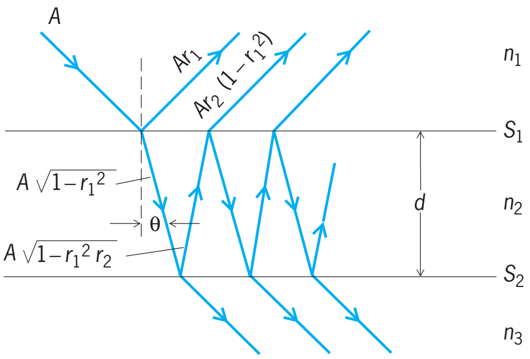


Fig. 9. Multiple reflection of wave between two surfaces.

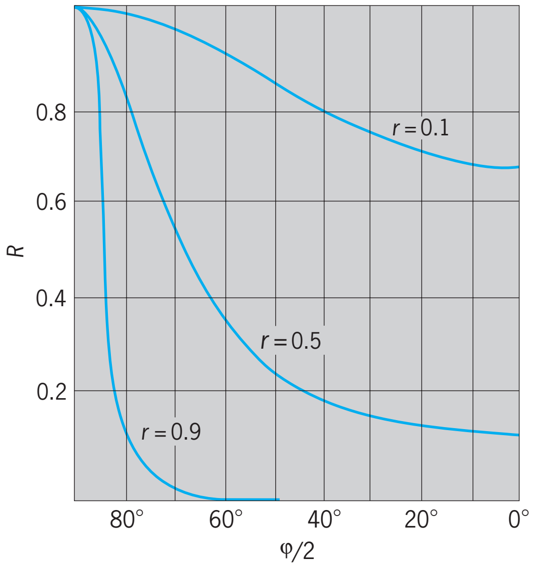


Fig. 10. The shape of multiple-beam fringes for different values of surface reflectivity.

Fig. 9, the different waves at a plane somewhere above S_1 can be represented by the following expressions:

$$\begin{aligned} &\text{Incoming wave} && = A \sin \omega t \\ &\text{First reflected wave} && = Ar_1 \sin \omega t \\ &\text{Second reflected wave} && = A(1 - {r_1}^2)r_2 \sin(\omega t + \varphi) \\ &\text{Third reflected wave} && = -A(1 - {r_1}^2)^2 r_1 {r_2}^2 \\ & && \qquad \times \sin(\omega t + 2\varphi) \end{aligned}$$

By inspection of these terms, one can write down the complete series. As in Eq. (3), the sine terms can be broken down and coefficients collected. A simpler method is to multiply each term by $i = \sqrt{-1}$ and add a cosine term with the same coefficient and argument. The individual terms then are all of the form of expression (39), where m is an integer.

$$Be^{-i\omega t}e^{-im\varphi} \tag{39}$$

The individual terms of expression (39) can be easily summed. For the reflected wave one obtains Eq. (40). Again, as in the two-beam case, the minimum in the reflectivity R is obtained when $\varphi = N\pi$, where N is an odd integer and $r_1 = r_2$. The fringe shape, however, can be quite different from the earlier case, depending on the values of the reflectives r_1 and r_2. The greater these values, the sharper become the fringes.

$$R = \frac{r_1 + r_2 e^{-i\varphi}}{1 + r_1 r_2 e^{-i\varphi}} \tag{40}$$

It was shown earlier how two-beam interference could be used to measure the contour of a surface. In this technique, a flat glass test plate was placed over the surface to be examined and monochromatic interference fringes were formed between the test surface and the surface of the plate. These two-beam fringes have intensities which vary as the cosine squared of the path difference. It is very difficult with such fringe to detect variations in fringe straightness or, in other terms, variations of surface planarity that are smaller than 1/20 wavelength. If the surface to be examined is coated with silver and the test surface is also coated with a partially transmitting metallic coat, the reflectivity increases to a point where many beams are involved in the formation of the interference fringes. The shape of the fringes is given by Eq. (40). The shape of fringes for different values of r is shown in **Fig. 10**. With high-reflectivity fringes, the sensitivity to a departure from planarity is increased far beyond 1/20 wavelength.

It is thus possible with partially silvered surfaces to get a much better picture of small irregularities than with uncoated surfaces. The increase in sensitivity is such that steps in cleaved mica as small as 1 nm in height can be seen by examining the monochromatic interference fringes produced between a silvered mica surface and a partially silvered glass flat. *See* INTERFERENCE FILTERS; INTERFEROMETRY.

Bruce H. Billings

Bibliography. M. Born and E. Wolf, *Principles of Optics*, 7th ed., 1999; F. A. Jenkins and H. E. White, *Fundamentals of Optics*, 4th ed., 1976; J. R. Meyer-Arendt, *Introduction to Classical and Modern Optics*, 4th ed., 1995; F. L. Pedrotti and L. S. Pedrotti, *Introduction to Optics*, 3d ed., 2006.

Interferometry

The design and use of optical interferometers. Optical interferometers based on both two-beam interference and multiple-beam interference of light are extremely powerful tools for metrology and spectroscopy. A wide variety of measurements can be performed, ranging from determining the shape of a surface to an accuracy of less than a millionth of an inch (25 nanometers) to determining the separation, by millions of miles, of binary stars. In spectroscopy, interferometry can be used to determine the hyperfine structure of spectrum lines. By using lasers in classical interferometers as well as holographic interferometers and speckle interferometers, it is possible to perform deformation, vibration, and contour measurements of diffuse objects that could not previously be performed.

Basic classes of interferometers. There are two basic classes of interferometers: division of wavefront and division of amplitude. **Figure 1** shows two arrangements for obtaining division of wavefront. For the Young's double pinhole interferometer (Fig. 1*a*), the light from a point source illuminates two pinholes. The light diffracted by these pinholes gives the interference of two point sources. For the Lloyd's mirror experiment (Fig. 1*b*), a mirror is used to provide a second image S_2 of the point source S_1, and in the region of overlap of the two beams the interference of two spherical beams can be observed. There are many other ways of obtaining division of wavefront; however, in each case the light leaving the source is spatially split, and then by use of diffraction, mirrors, prisms, or lenses the two spatially separated beams are superimposed.

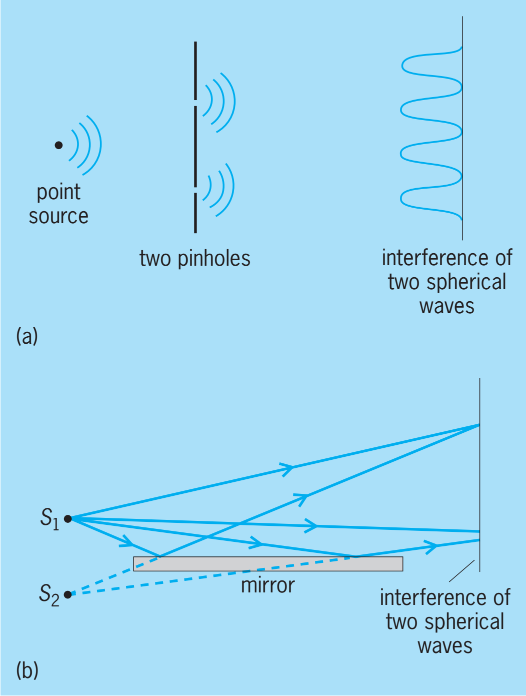


Fig. 1. **Interference produced by division of wavefront. (*a*) Young's two-pinhole interferometer. (*b*) Lloyd's mirror.**

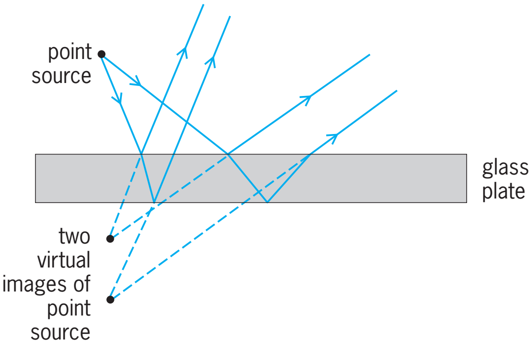


Fig. 2. **Division of amplitude.**

Figure 2 shows one technique for obtaining division of amplitude. For division-of-amplitude interferometers a beam splitter of some type is used to pick off a portion of the amplitude of the radiation which is then combined with a second portion of the amplitude. The visibility of the resulting interference fringes is a maximum when the amplitudes of the two interfering beams are equal. *See* INTERFERENCE OF WAVES.

Michelson interferometer. The Michelson interferometer (**Fig. 3**) is based on division of amplitude. Light from an extended source *S* is incident on a partially reflecting plate (beam splitter) P_1. The light transmitted through P_1 reflects off mirror M_1 back to plate P_1. The light which is reflected proceeds to M_2 which reflects it back to P_1. At P_1, the two waves are again partially reflected and partially transmitted, and a portion of each wave proceeds to the receiver *R*, which may be a screen, a photocell, or a human eye. Depending on the difference between the distances from the beam splitter to the mirrors M_1 and M_2, the two beams will interfere constructively or destructively. Plate P_2 compensates for the thickness of P_1. Often when a quasimonochromatic light source is used with the interferometer, compensating plate P_2 is omitted.

The function of the beam splitter is to superimpose (image) one mirror onto the other. When the mirrors' images are completely parallel, the interference fringes appear circular. If the mirrors are slightly inclined about a vertical axis, vertical fringes are formed across the field of view. These fringes can be formed in white light if the path difference in part of the field of view is made zero. Just as in other interference experiments, only a few fringes will appear in white light, because the difference in path will be different for wavelengths of different colors. Accordingly, the fringes will appear colored close to zero path difference, and will disappear at larger path differences where the fringe maxima and minima for the different wavelengths overlap. If light reflected off the beam splitter experiences a one-half-cycle relative phase shift, the fringe of zero path difference is black, and can be easily distinguished from the neighboring fringes. This makes use of the instrument relatively easy.

The Michelson interferometer can be used as a spectroscope. Consider first the case of two close spectrum lines as a light source for the instrument. As the mirror M_1 is shifted, fringes from each spectral line will cross the field. At certain path differences between M_1 and M_2, the fringes for the two spectral lines will be out of phase and will essentially disappear; at other points they will be in phase and will be reinforced. By measuring the distance between successive maxima in fringe contrast, it is possible to determine the wavelength difference between the lines.

This is a simple illustration of a very broad use for any two-beam interferometer. As the path length *L* is changed, the variation in intensity *I*(*L*) of the light coming from an interferometer gives information on the basis of which the spectrum of the input light can be derived. The equation for the intensity of the emergent energy can be written as

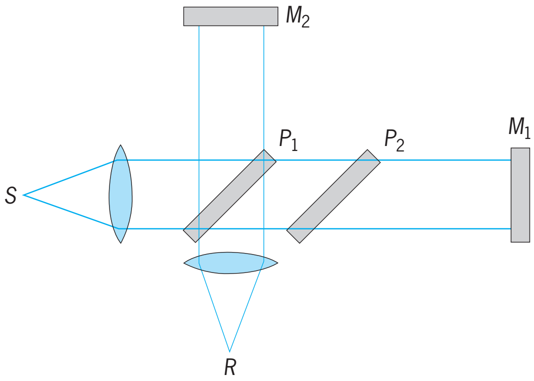


Fig. 3. **Michelson interferometer.**

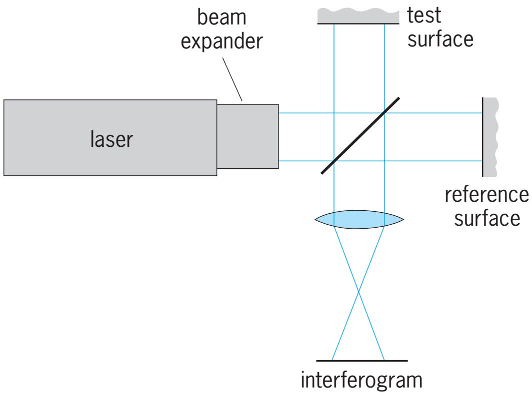


Fig. 4. Twyman-Green interferometer for testing flat surfaces.

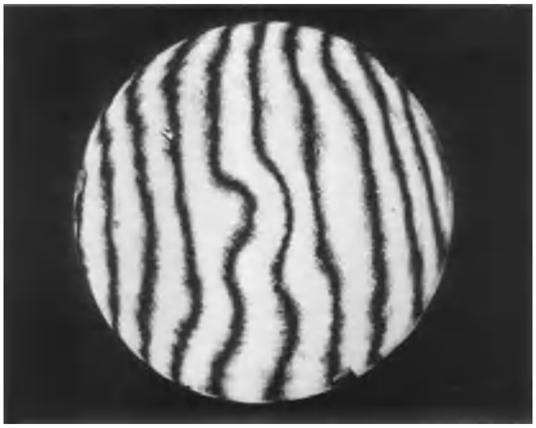


Fig. 5. Interferogram obtained with the use of a Twyman-Green interferometer to test a flat surface.

$$I(L) = \int_0^\infty I(\lambda)\cos^2\left(\frac{\beta L}{\lambda}\right) d\lambda \qquad (1)$$

Eq. (1), where β is a constant, and $I(\lambda)$ is the intensity of the incident light at different wavelengths λ. This equation applies when the mirror M_1 is moved linearly with time from the position where the path difference with M_2 is zero, to a position which depends on the longest wavelength in the spectrum to be examined. From Eq. (1), it is possible mathematically to recover the spectrum $I(\lambda)$. In certain situations, such as in the infrared beyond the wavelength region of 1.5 micrometers, this technique offers a large advantage over conventional spectroscopy in that its utilization of light is extremely efficient. *See* INFRARED SPECTROSCOPY.

Twyman-Green interferometer. If the Michelson interferometer is used with a point source instead of an extended source, it is called a Twyman-Green interferometer. The use of the laser as the light source for the Twyman-Green interferometer has made it an extremely useful instrument for testing optical components. The great advantage of a laser source is that it makes it possible to obtain bright, good-contrast, interference fringes even if the path lengths for the two arms of the interferometer are quite different. *See* LASER.

Figure 4 shows a Twyman-Green interferometer for testing a flat mirror. The laser beam is expanded to match the size of the sample being tested. Part of the laser light is transmitted to the reference surface, and part is reflected by the beam splitter to the flat surface being tested. Both beams are reflected back to the beam splitter, where they are combined to form interference fringes. An imaging lens projects the surface under test onto the observation plane.

Fringes (**Fig. 5**) show defects in the surface being tested. If the surface is perfectly flat, then straight, equally spaced fringes are obtained. Departure from the straight, equally spaced condition shows directly how the surface differs from being perfectly flat. For a given fringe, the difference in optical path between light going from laser to reference surface to observation plane and the light going from laser to test surface to observation plane is a constant. (The optical path is equal to the product of the geometrical path times the refractive index.) Between adjacent fringes (Fig. 5), the optical path difference changes by one wavelength, which for a helium-neon laser corresponds to 633 nm. The number of straight, equally spaced fringes and their orientation depend upon the tip-tilt of the reference mirror. That is, by tipping or tilting the reference mirror the difference in optical path can be made to vary linearly with distance across the laser beam.

Deviations from flatness of the test mirror also cause optical path variations. A height change of half a wavelength will cause an optical path change of one wavelength and a deviation from fringe straightness of one fringe. Thus, the fringes give surface height information, just as a topographical map gives height or contour information.

The existence of the essentially straight fringes provides a means of measuring surface contours relative to a tilted plane. This tilt is generally introduced to indicate the sign of the surface error, that is, whether the errors correspond to a hill or a valley. One way to get this sign information is to push in on the piece being tested when it is in the interferometer. If the fringes move toward the right when the test piece is pushed toward the beam splitter, then fringe deviations from straightness toward the right correspond to high points (hills) on the test surface and deviations to the left correspond to low points (valleys).

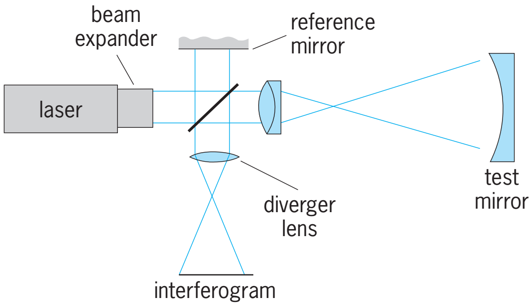


Fig. 6. Twyman-Green interferometer for testing spherical mirrors or lenses.

The basic Twyman-Green interferometer (Fig. 4) can be modified (**Fig. 6**) to test concave-spherical mirrors. In the interferometer, the center of curvature of the surface under test is placed at the focus of a high-quality diverger lens so that the wavefront is reflected back onto itself. After this retroflected wavefront passes through the diverger lens, it will be essentially a plane wave, which, when it interferes with the plane reference wave, will give interference fringes similar to those shown in Fig. 5 for testing flat surfaces. In this case it indicates how the concave-spherical mirror differs from the desired shape. Likewise, a convex-spherical mirror can be tested. Also, if a high-quality spherical mirror is used, the high-quality diverger lens can be replaced with the lens to be tested.

Fizeau interferometer. One of the most commonly used interferometers in optical metrology is the Fizeau interferometer, which can be thought of as a folded Twyman-Green interoferometer. In the Fizeau, the two surfaces being compared, which can be flat, spherical, or aspherical, are placed in close contact. The light reflected off these two surfaces produces interference fringes. For each fringe, the separation between the two surfaces is a constant. If the two surfaces match, straight, equally spaced fringes result. Surface height variations between the two surfaces cause the fringes to deviate from straightness or equal separations, where one fringe deviation from straightness corresponds to a variation in separation between the two surfaces by an amount equal to one-half of the wavelength of the light source used in the interferometer. The wavelength of a helium source, which is often used in a Fizeau interferometer, is 587.56 nm; hence one fringe corresponds to a height variation of approximately 0.3 μm.

Mach-Zehnder interferometer. The Mach-Zehnder interferometer (**Fig. 7**) is a variation of the Michelson interferometer and, like the Michelson interferometer, depends on amplitude splitting of the wavefront. Light enters the instrument and is reflected and transmitted by the semitransparent mirror M_1. The reflected portion proceeds to M_3, where it is reflected through the cell C_2 to the semitransparent mirror M_4. Here it combines with the light transmitted by M_1 to produce interference. The light transmitted by M_1 passes through a cell C_1, which is similar to C_2 and is used to compensate for the windows of C_1.

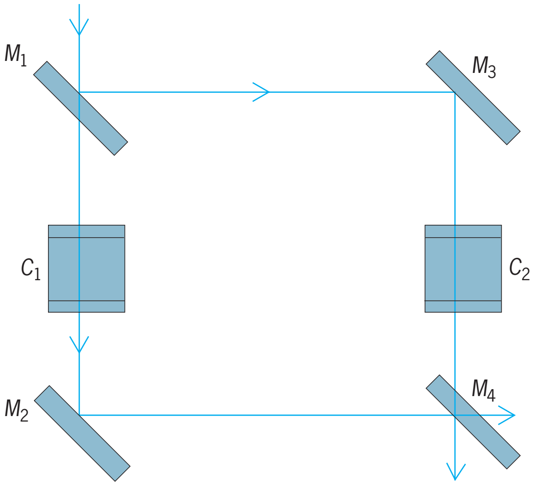


Fig. 7. Mach-Zehnder interferometer.

The major application of this instrument is in studying airflow around models of aircraft, missiles, or projectiles. The object and associated airstream are placed in one arm of the interferometer. Because the air pressure varies as it flows over the model, the index of refraction varies, and thus the effective path length of the light in this beam is a function of position. When the variation is an odd number of half-waves, the light will interfere destructively and a dark fringe will appear in the field of view. From a photograph of the fringes, the flow pattern can be mathematically derived.

A major difference between the Mach-Zehnder and the Michelson interferometer is that in the Mach-Zehnder the light goes through each path in the instrument only once, whereas in the Michelson the light traverses each path twice. This double traversal makes the Michelson interferometer extremely difficult to use in applications where spatial location of index variations is desired. The incoming and outgoing beams tend to travel over slightly different paths, and this lowers the resolution because of the index gradient across the field.

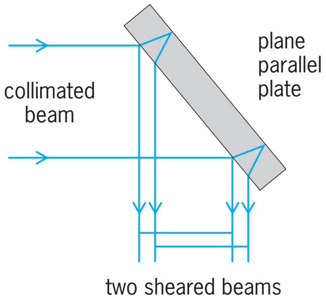


Fig. 8. Lateral shear interferometer.

Shearing interferometers. In a lateral-shear interferometer, an example of which is shown in **Fig. 8**, a wavefront is interfered with a shifted version of itself. A bright fringe is obtained at the points where the slope of the wavefront times the shift between the two wavefronts is equal to an integer number of wavelengths. That is, for a given fringe the slope or derivative of the wavefront is a constant. For this reason a lateral-shear interferometer is often called a differential interferometer.

Another type of shearing interferometer is a radial-shear interferometer. Here, a wavefront is interfered with an expanded version of itself. This interferometer is sensitive to radial slopes.

The advantages of shearing interferometers are that they are relatively simple and inexpensive, and since the reference wavefront is self-generated, an external wavefront is not needed. Since an external reference beam is not required, the source

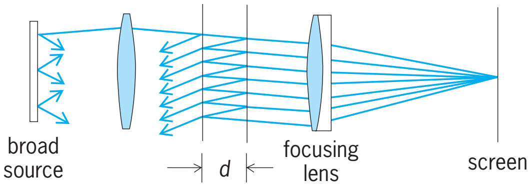


Fig. 9. Fabry-Perot interferometer.

requirements are reduced from those of an interferometer such as a Twyman-Green. For this reason, shearing interferometers, in particular lateral-shear interferometers, are finding much use in applications such as adaptive optics systems for correction of atmospheric turbulence where the light source has to be a star, or planet, or perhaps just reflected sunlight. *See* ADAPTIVE OPTICS.

Michelson stellar interferometer. A Michelson stellar interferometer can be used to measure the diameter of stars which are as small as 0.01 second of arc. This task is impossible with a ground-based optical telescope since the atmosphere limits the resolution of the largest telescope to not much better than 1 second of arc.

The Michelson stellar interferometer is a simple adaptation of Young's two-slit experiment. In its first form, two slits were placed over the aperture of a telescope. If the object being observed were a true point source, the image would be crossed with a set of interference bands. A second point source separated by a small angle from the first would produce a second set of fringes. At certain values of this angle, the bright fringes in one set will coincide with the dark fringes in the second set. The smallest angle α at which the coincidence occurs will be that angle subtended at the slits by the separation of the peak of the central bright fringe from the nearest dark fringe. This angle is given by Eq. (2), where d is the

$$\frac{\lambda}{2d} = \alpha \tag{2}$$

separation of the slits, λ the dominant wavelength of the two sources, and α their angular separation. The measurement of the separation of the sources is performed by adjusting the separation d between the slits until the fringes vanish.

Consider now a single source in the shape of a slit of finite width. If the slit subtends an angle at the telescope aperture which is larger than α, the interference fringes will be reduced in contrast. For various line elements at one side of the slit, there will be elements of angle α which will cancel the fringes from the first element. By induction, it is clear that for a separation d' such that the slit source subtends an angle as given by Eq. (3) the fringes from a single

$$\alpha' = \frac{\lambda}{d'} \tag{3}$$

slit will vanish completely. For additional information on the Michelson stellar interferometer *see* DIFFRACTION

Fabry-Perot interferometer. All the interferometers discussed above are two-beam interferometers. The Fabry-Perot interferometer (**Fig. 9**) is a multiple-beam interferometer since the two glass plates are partially silvered on the inner surfaces, and the incoming wave is multiply reflected between the two surfaces. The position of the fringe maxima is the same for multiple-beam interference as two-beam interference; however, as the reflectivity of the two surfaces increases and the number of interfering beams increases, the fringes become sharper.

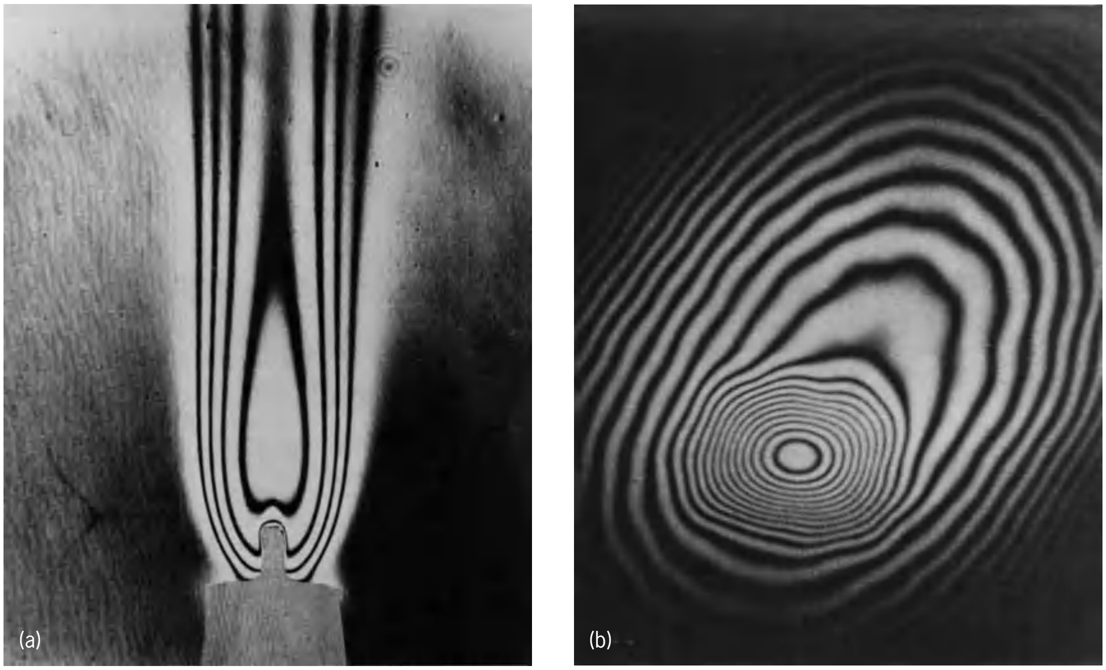


Fig. 10. Double-exposure holographic interferograms. (*a*) Interferogram of candle flame. (*b*) Interferogram of debanded region of honeycomb construction panel. (*From C. M. Vest*, *Holographic Interferometry*, *John Wiley and Sons*, *1979*)

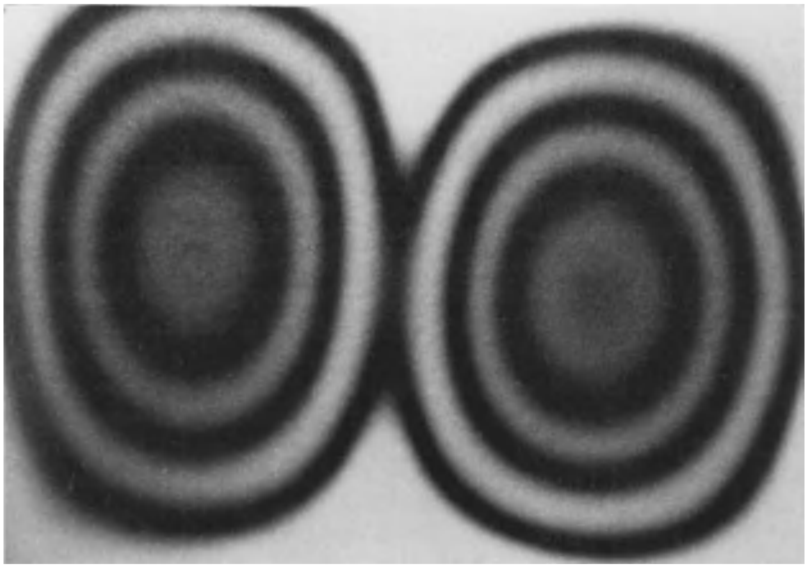


Fig. 11. Photograph of time-average holographic interferogram. (*From C. M. Vest, Holographic Interferometry, John Wiley and Sons, 1979*)

A quantity of particular interest in a Fabry-Perot is the ratio of the separation of adjacent maxima to the half-width of the fringes. It can be shown that this ratio, known as the finesse, is given by Eq. (4), where

$$\mathscr{F} = \frac{\pi\sqrt{R}}{1-R} \tag{4}$$

R is the reflectivity of the silvered surfaces.

The multiple-beam Fabry-Perot interferometer is of considerable importance in modern optics for spectroscopy. All the light rays incident on the Fabry-Perot at a given angle will result in a single circular fringe of uniform irradiance. With a broad diffuse source, the interference fringes will be narrow concentric rings, corresponding to the multiple-beam transmission pattern. The position of the fringes depends upon the wavelength. That is, each wavelength gives a separate fringe pattern. The minimum resolvable wavelength difference is determined by the ability to resolve close fringes. The ratio of the wavelength λ to the least resolvable wavelength difference Δλ is known as the chromatic resolving power $\mathscr{R}$. At nearly normal incidence it is given by Eq. (5), where *n* is the refractive index between the

$$\mathscr{R} = \frac{\lambda}{(\Delta\lambda)_{\min}} = \mathscr{F}\,\frac{2nd}{\lambda} \tag{5}$$

two mirrors separated a distance *d*. For a wavelength of 500 nm, $nd = 10$ mm, and $R = 90\%$, the resolving power is well over 10^6. *See* REFLECTION OF ELECTROMAGNETIC RADIATION; RESOLVING POWER (OPTICS).

When Fabry-Perot interferometers are used with lasers, they are generally used in the central spot scanning mode. The interferometer is illuminated with a collimated laser beam, and all the light transmitted through the Fabry-Perot is focused onto a detector, whose output is displayed on an oscilloscope. Often one of the mirrors is on a piezoelectric mirror mount. As the voltage to the piezoelectric crystal is varied, the mirror separation is varied. The light output as a function of mirror separation gives the spectral frequency content of the laser source.

Holographic interferometry. A wave recorded in a hologram is effectively stored for future reconstruction and use. Holographic interferometry is concerned with the formation and interpretation of the fringe pattern which appears when a wave, generated at some earlier time and stored in a hologram, is later reconstructed and caused to interfere with a comparison wave. It is the storage or time-delay aspect which gives the holographic method a unique advantage over conventional optical interferometry.

A hologram can be made of an arbitrarily shaped, rough scattering surface, and after suitable processing, if the hologram is illuminated with the same reference wavefront used in recording the hologram, the hologram will produce the original object wavefront. If the hologram is placed back into its original position, a person looking through the hologram will see both the original object and the image of the object stored in the hologram. If the object is now slightly deformed, interference fringes will be produced which tell how much the surface is deformed. Between adjacent fringes the optical path between the source and viewer has changed by one wavelength. While the actual shape of the object is not determined, the change in the shape of the object is measured to within a small fraction of a wavelength, even though the object's surface is rough compared to the wavelength of light.

Double-exposure. Double-exposure holographic interferometry (**Fig. 10**) is similar to real-time holographic interferometry described above, except now two exposures are made before processing: one exposure with the object in the undeformed state and a second exposure after deformation. When the hologram reconstruction is viewed, interference fringes

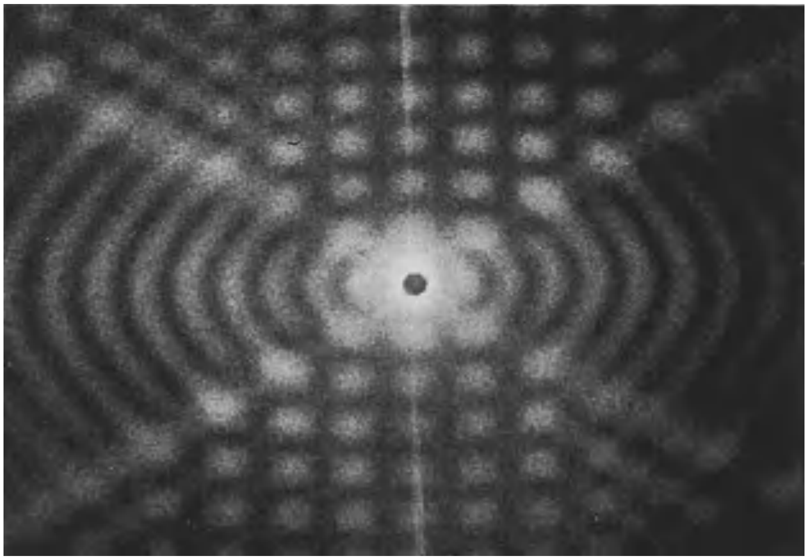


Fig. 12. Diffraction patterns from time-averaged speckle interferogram of a surface vibrating in its own plane with a figure-eight motion. (*From J. C. Dainty, Laser Speckle and Related Phenomena, Springer-Verlag, 1975*)

will be seen which show how much the object was deformed between exposures.

The advantage of the double-exposure technique over the real-time technique is that there is no critical replacement of the hologram after processing. The disadvantage is that continuous comparison of surface displacement relative to an initial state cannot be made, but rather only the difference between two states is determined.

Time-average. In time-average holographic interferometry (**Fig. 11**) a time-average hologram of a vibrating surface is recorded. If the maximum amplitude of the vibration is limited to some tens of light wavelengths, illumination of the hologram yields an image of the surface on which is superimposed several interference fringes which are contour lines of equal displacement of the surface. Time-average holography enables the vibrational amplitudes of diffusely reflecting surfaces to be measured with interferometric precision. *See* HOLOGRAPHY.

Speckle interferometry. A random intensity distribution, called a speckle pattern, is generated when light from a highly coherent source, such as a laser, is scattered by a rough surface. The use of speckle patterns in the study of object displacements, vibration, and distortion is becoming of more importance in the nondestructive testing of mechanical components. For example, time-averaged speckle photographs can be used to analyze the vibrations of an object in its plane. In physical terms the speckles in the image are drawn out into a line as the surface vibrates, instead of being double as in the double-exposure technique. The diffraction pattern of this smeared-out speckle-pattern recording is related to the relative time spent by the speckle at each point of its trajectory (**Fig. 12**). *See* NONDESTRUCTIVE EVALUATION.

Speckle interferometry can be used to perform astronomical measurements similar to those performed by the Michelson stellar interferometer. Stellar speckle interferometry is a technique for obtaining diffraction-limited resolution of stellar objects despite the presence of the turbulent atmosphere that limits the resolution of ground-based telescopes to approximately 1 second of arc. For example, the diffraction limit of the 200-in.-diameter (5-m) Palomar Mountain telescope is approximately 0.02 second of arc, ${}^{1}/_{\sqrt{50}}$ the resolution limit set by the atmosphere.

The first step of the process is to take a large number, perhaps 100, of short exposures of the object, where each photo is taken for a different realization of the atmosphere. Next the optical diffraction pattern, that is, the squared modulus of the Fourier transform of all the short-exposure photographs, is added. By taking a further Fourier transform of each ensemble average diffraction pattern, the ensemble average of the spatial autocorrelation of the diffraction-limited images of each object is obtained. *See* SPECKLE.

Phase-shifting interferometry. Electronic phase-measurement techniques can be used in interferometers such as the Twyman-Green, where the phase distribution across the interferogram is being measured. Phase-shifting interferometry is often used for these measurements since it provides for rapid precise measurement of the phase distribution. In phase-shifting interferometry, the phase of the reference beam in the interferometer is made to vary in a known manner. This can be achieved, for example, by mounting the reference mirror on a piezoelectric transducer. By varying the voltage on the transducer, the reference mirror is moved a known amount to change the phase of the reference beam a known amount. A solid-state detector array is used to detect the intensity distribution across the interference pattern. This intensity distribution is read into computer memory three or more times, and between each intensity measurement the phase of the reference beam is changed a known amount. From these three or more intensity measurements, the phase across the interference pattern can be determined to within a fraction of a degree. James C. Wyant

Bibliography. P. Hariharan, *Basics of Interferometry*, 1991; E. Hecht and A. Zajac, *Optics*, 3d ed., 1997; G. Hernandez, *Fabry-Perot Interferometers*, 1988; D. Malacara (ed.), *Optical Shop Testing*, 2d ed., 1992; P. K. Rastogi, *Holographic Interferometry: Principles and Methods*, 1994; R. S. Sirohi (ed.), *Speckle Metrology*, 1993; W. H. Steel, *Interferometry*, 2d ed., 1985.

Interhalogen compounds

The elements of the halogen family (fluorine, chlorine, bromine, and iodine) possess an ability to react with each other to form a series of binary interhalogen compounds (or halogen halides) of general composition given by XY; where *n* can have the values 1, 3, 5, and 7, and where X is the heavier (less electronegative) of the two elements. All possible diatomic compounds of the first four halogens have been prepared. In other groups a varying number of possible combinations is absent. Although attempts have been made to prepare ternary interhalogens, they have been unsuccessful; there is considerable doubt that such compounds can exist. *See* HALOGEN ELEMENTS.

Formation. In general, interhalogen compounds are formed when the free halogens are mixed as gases, or, as in the case of iodine chlorides and bromides, by reacting solid iodine with liquid chlorine or bromine. Most of the nonfluorinated interhalogens also readily form when solutions of the halogens in inert solvent (for example, carbon tetrachloride) are mixed. It is also possible to form them by the reaction of a halogen with a salt of a more electropositive halogen, such as $KI + Cl_2 \rightarrow KCl + ICl$. Higher polyhalides can also be prepared by reacting more electronegative halogen with a corresponding halogen halide, for example, $ICl + Cl_2 \rightarrow ICl_3$ or $ClF_3 + F_2 \rightarrow ClF_5$. Chlorine pentafluoride can also be prepared by reacting a $MClF_4$ salt with fluorine ($M =$ alkali metal), $MClF_4 + F_2 \rightarrow MF + ClF_5$. A list of known interhalogen compounds and some of their

TABLE 1. Known interhalogen compounds

	XY	XY_3	XY_5	XY_7
	ClF	ClF_3	ClF_5	IF_7
mp	−154°C (−245°F)	−76°C (−105°F)	−103°C (−153°F)	
bp	−101°C (−150°F)	12°C (54°F)	−14°C (6.8°F)	
				4.77°C (40.6°F) (sublimes)
	BrF	BrF_3	BrF_5	
mp	≈−33°C (91°F)	8.77°C (47.8°F)	−62.5°C (−80.5°F)	
bp	≈20°C (68°F)	125°C (257°F)	40.3°C (105°F)	
	IF	IF_3	IF_5	
mp	—	−28°C (−18°F)	10°C (50°F)	
bp	—	—	101°C (214°F)	
	BrCl	ICl_3*		
mp	≈−54°C (−65°F)	101°C (214°F)		
bp	—	—		
	ICl†			
mp	27.2 C(α)			
bp	≈100°C (212°F)			
	IBr			
mp	40°C (104°F)			
bp	119°C (246°F)			

*In the solid state the compound forms a dimer.

†Unstable β-modification exists, mp 14° C (57° F).

physical properties is given in **Table 1**. Interhalogen compounds containing astatine have not been isolated as yet, although the existence of AtI and AtBr has been demonstrated by indirect measurements.

Stability. Thermodynamic stability of the interhalogen compounds varies within rather large limits. In general, for a given group the stability increases with increasing difference in electronegativity between the two halogens. Thus for XY group the free energy of formation of the interhalogens, relative to the elements in their standard conditions, falls in the following order: IF > BrF > ClF > ICl > IBr > BrCl. It should be noted, however, that the fluorides of this series can be obtained only in minute quantities since they readily undergo disproportionation reaction, for example, $5IF \rightarrow 2I_2 + IF_5$. The least-stable compound, bromine chloride, has only recently been isolated in the pure state. Decrease of stability with decreasing difference of electronegativity is readily apparent in the higher interhalogens since compounds such as $BrCl_3$, IBr_3, or ICl_5 are unknown. The only unambiguously prepared interhalogen containing eight halogen atoms is IF_7. *See* ELECTRONEGATIVITY.

Reactivity. The reactivity of the polyhalides reflects the reactivity of the halogens they contain. In general, they behave as strong oxidizing and halogenating agents. Most halogen halides (especially halogen fluorides) readily attack metals, yielding the corresponding halide of the more electronegative halogen. In the case of halogen fluorides, the reaction results in the formation of the fluoride, in which the metal is often found in its highest oxidation state, for example, AgF_2, CoF_3, and so on. Noble metals, such as platinum, are resistant to the attack of the interhalogens at room temperature. Halogen fluorides are often handled in nickel vessels, but in this case the resistance to attack is due to the formation of a protective layer of nickel(II) fluoride.

All halogen halides readily react with water. Such reactions can be quite violent and, with halogen fluorides, they may be explosive. Reaction products vary depending on the nature of the interhalogen compound. For example, in the case of chlorine trifluoride, reaction with an excess of water yields HF, Cl_2, and O_2 as reaction products.

The same reactivity is observed with organic compounds. The nonfluorinated interhalogens do not react with completely halogenated hydrocarbons, and solutions of ICl, IBr, and BrCl are quite stable in dry carbon tetrachloride or hexachloroethane, as well as in fluorocarbons, as long as the solvents are very dry. They readily react with aliphatic and aromatic hydrocarbons and with oxygen- or nitrogen-containing compounds. The reaction rates, however, can be rather slow and dilute solutions of ICl can be stable for several hours in solvents such as nitrobenzene. Halogen fluorides usually react vigorously with chlorinated hydrocarbons, although IF_5 can be dissolved in carbon tetrachloride and BrF_3 can be dissolved in Freon 113 without decomposition. All halogen fluorides react explosively with easily oxidizable organic compunds. *See* FLUORINE.

Halogen halides, like halogens, act as Lewis acids and under proper experimental conditions may form a series of stable complexes with various organic electron donors. For example, mixing of carbon tetrachloride solutions of pyridine and of iodine monochloride leads to the formation of a solid complex, $C_5H_5 \cdot ICl$. The same reaction can occur with other heterocyclic amines and ICl, IBr, or ICl_3. Addition compounds of organic electron donors with IF, IF_3, and IF_5 have been reported. In all cases it is the iodine atom which is directly attached to the donor atom.

A number of interhalogen compounds conduct electrical current in the liquid state. Among these are ICl and BrF_3. For example, electrical conductance of molten iodine monochloride is comparable to a concentrated aqueous solution of a strong electrolyte (4.52×10^{-3} ohm^{-1} cm^{-1} at 30.6°C or 87.08°F). The conductances, however, are much smaller than those of fused salts and, therefore, it can be concluded that the bonding in these compounds is largely covalent. Electrical conductance is due to self-ionization reactions, as shown in reactions (1) and (2).

$$3ICl \rightleftharpoons I_2Cl^+ + ICl_2^- \tag{1}$$

$$2BrF_3 \rightleftharpoons BrF_2^+ + BrF_4^- \tag{2}$$

The above behavior leads to the possibility of studying acid-base reactions. In these systems an acid is any compound which generates the solvo-cation, while a base would generate solvo-anions. Thus SbF_5 would be an acid in liquid bromine trifluoride, reaction (3), while an electrovalent fluoride would be a base, reaction (4).

$$SbF_5 + BrF_3 \rightleftharpoons BrF_2^+ + SbF_6^- \tag{3}$$

$$KF + BrF_3 \rightleftharpoons K^+ + BrF_4^- \tag{4}$$

See SUPERACID.

Analogy with acid-base reactions in water is obvious, as shown by reactions (5) and (6). Such relations

$$K^+OH^- + H_3O^+Cl^- \rightarrow K^+Cl^- + 2H_2O \tag{5}$$

$$K^+BrF_4^- + BrF_2^+SbF_6^- \rightarrow KSbF_6 + 2BrF_3 \tag{6}$$

have been studied in BrF_3, ClF_3, and IF_5. Numerous salts of the interhalogen acid-base systems have been isolated and studied.

Thus, numerous compounds have been formed containing either interhalogen anions (solvo-anions) or interhalogen cations (solvo-cations) simply by adding the appropriate acid or base to a liquid halogen halide or a halogen halide in an appropriate nonaqueous solvent. In addition, cations derived from previously unknown compounds can be prepared by using powerful oxidizing agents, such as KrF^+ salts. For example, even though to date, a compound containing BrF^{+}_{6} has been prepared according to reaction (7).

$$BrF_5 + KrF^+AsF_6^- \rightarrow BrF_6^+AsF_6^- + Kr \tag{7}$$

TABLE 2. Known interhalogen anions

Three-membered	Five-membered	Seven-membered	Nine-membered
ClF_2^-	ClF_4^-	ClF_6^-	IF_3^-
BrF_2^-	BrF_4^-	BrF_6^-	
ICl_2^-	IF_4^-	IF_6^-	
IBr_2^-	ICl_4^-		
$IBrCl^-$	$I_2Cl_3^-$		
$BrCl_2^-$	$I_2Cl_2Br^-$		
I_2Cl^-	$I_2ClBr_2^-$		
Br_2Cl^-	I_4Cl^-		
I_2Br^-			

TABLE 3. Known interhalogen cations

Three-membered	Five-membered	Seven-membered
Cl_2F^+	ClF_4^+	ClF_6^+
ClF_2^+	BrF_4^+	BrF_6^+
BrF_2^+	IF_4^+	
ICl_2^+		
I_2Cl^+		
IBr_2^+		
I_2Br^+		
$BrCl_2^+$		
Br_2Cl^+		
$IBrCl^+$		

Pentahalides can also be formed by the addition of an interhalogen compound to a trihalide ion as shown in reaction (8).

$$ICl + ICl_2^- \rightarrow I_2Cl_3^- \tag{8}$$

A compilation of interhalogen anions and cations which have been previously prepared is given in **Tables 2** and **3**.

Terry Surles

Bibliography. V. Gutmann (ed.), *MTP International Review of Science: Inorganic Chemistry*, Ser. 1, vol. 3, 1975.

Intermediate-frequency amplifier

An amplifying circuit in a radio-frequency (RF) receiver that processes and enhances a downconverted or modulated signal. Signal frequency spectrum downconversion is achieved by multiplying the radio-frequency signal by a local oscillator signal in a circuit known as a mixer. This multiplication produces two signals whose frequency content lies about the sum and difference frequencies of the center frequency of the original signal and the oscillator frequency. A variable local oscillator is used in the receiver to hold the difference-signal center frequency constant as the receiver is tuned. The constant frequency of the downconverted signal is called the intermediate frequency (IF), and it is this signal that is processed by the intermediate-frequency amplifier.

Unfortunately, radio-frequency signals both higher and lower than the local oscillator frequency by a difference equal to the intermediate frequency will produce the intermediate frequency. One of these is the desired signal; the undesired signal is called an image. *See* MIXER; OSCILLATOR.

Superheterodyne receiver. Aside from demodulation and conversion, the purpose of each stage of a radio receiver is to improve the signal-to-noise ratio (SNR) through a combination of signal amplification and noise/interference suppression. The first stage of a superheterodyne receiver is generally a broadband tuned radio-frequency amplifier intended to improve the signal-to-noise ratio of a selected signal whose frequency lies anywhere within the receiver's tuning range. The second stage is a variable-frequency mixer or downconverter that shifts the frequency of the signal of interest to the predetermined intermediate frequency by use of a local oscillator (LO). Waveform impurities in the LO signal and nonlinearities in the RF amplifier and mixer may interact with the received signal to generate undesirable false signals that are replicas or images of the received signal but displaced in frequency. The third stage of that receiver is a high-quality intermediate-frequency amplifier that has been optimized to improve the quality of the signal in a very narrow frequency band about the fixed intermediate frequency by greatly attenuating out-of-band noise, adjacent-channel signals, and images. More sophisticated digital intermediate-frequency stages can also reduce broadband interference.

Unlike the broadband tunable radio-frequency amplifier, the intermediate-frequency amplifier is designed to operate over a narrow band of frequencies centered about a dedicated fixed frequency (the intermediate frequency); therefore, the intermediate-frequency amplifier can be an extremely efficient stage. If the intermediate frequency is on the order of a few megahertz, the undesirable images may be efficiently rejected, but narrow-band filtering for noise and adjacent-channel-signal rejection is difficult and expensive because of the high ratio of the intermediate frequency to the bandwidth of the intermediate-frequency amplifier. If the intermediate frequency is much smaller, say, on the order of a few hundred kilohertz, then inexpensive and more selective filters are possible that can separate the desired signal from closely packed adjacent signals, but they do not reject images very well. A high-quality double-conversion receiver combines the best of both approaches by cascading both high- and low-frequency intermediate-frequency stages that are separated by a second fixed-frequency mixer. The last intermediate-frequency stage of a digital or analog superheterodyne receiver is followed by the detector, which extracts the information from the signal. *See* HETERODYNE PRINCIPLE; RADIO-FREQUENCY AMPLIFIER; RADIO RECEIVER; TELEVISION RECEIVER.

Applications. The superheterodyne structure is common for television, ground-based and satellite communications, cell phones, ground-based and airborne radar, navigation, and many other receivers. The intermediate-frequency amplifier function is ubiquitous.

Gain and bandwidth. The measure of quality of an intermediate-frequency amplifier is how well it forms a "window" in the frequency spectrum for observing signals of interest and rejecting all other signals (including images) and noise. Maximum gain is usually at the intermediate frequency. The intermediate-frequency-amplifier gain decreases at frequencies both higher and lower than the intermediate frequency. The difference between the upper and lower frequencies at which the signal power drops to one-half of the maximum signal power (usually at the intermediate frequency) is known as the selectivity or the 6-dB bandwidth. The difference between the upper and lower frequencies at which the signal power drops to one-thousandth of the maximum signal power is known as the 60-dB bandwidth. The ratio of the 60-dB bandwidth and the 6-dB bandwidth is known as the shape factor, a number that must always be greater than one. A small shape factor indicates good noise rejection because the filter rejects any signal outside of the bandwidth of interest. The best selectivity matches the bandwidth of the signal.

Because a single receiver may be used for several different applications, a different intermediate-frequency filter may be employed for each application. The filter requirements vary widely with application. For example, signal bandwidth for broadcast-quality frequency modulation (FM) may be in the tens of kilohertz; for communications-quality FM, 7–10 kHz; for broadcast-quality amplitude modulation (AM), 5 kHz; for communications-quality AM, 3 kHz; for single sideband (SSB), 1–2 kHz; and for continuous-wave (CW) communications (radio-telegraphy), under 100 Hz. *See* BANDWIDTH REQUIREMENTS (COMMUNICATIONS); GAIN; SELECTIVITY.

Analog IF amplifiers. Analog (or continuous-time) intermediate-frequency amplifiers consist of narrow-band tuned circuits that filter out nearly all signals and noise except the desired signal at the intermediate frequency, and amplification to increase the signal strength. The original tuned-circuit structures were made of capacitors and inductors. Today virtually all high-Q intermediate-frequency filters are only a part of a monolithic integrated circuit that contains the entire radio receiver, including fully integrated RC/op-amp active filters, piezoelectric crystal filters, magnetostrictive lines, and micro-electro-mechanical systems (MEMS). *See* ELECTRIC FILTER; INTEGRATED CIRCUITS; MAGNETOSTRICTION; MICRO-ELECTRO-MECHANICAL SYSTEMS (MEMS); OPERATIONAL AMPLIFIER; PIEZOELECTRICITY; Q (ELECTRICITY).

Digital IF amplifiers. This misnomer came to be because the concept is easy to grasp in view of its parallel to analog IF amplifiers. Either digital IF stage or digital IF filter would be a more appropriate name.

The current goal of radio-frequency system design is to make an analog-to-digital converter (ADC) directly connected to the antenna to permit realization of an all-digital radio receiver. Fast devices permit analog-to-digital converters that operate at near-terahertz (10^{12} Hz) speed. In 1999 such devices were still costly and power-hungry, but the same was true of bit-serial near-megahertz devices around 1970. There now exist reasonably priced and powered radio receivers that consist of an analog radio-frequency amplifier and first converter, followed by a high-speed analog-to-digital converter, which is connected to the remainder of the digital-receiver processing, beginning with the intermediate-frequency stage. The digital-receiver function can be produced as a dedicated integrated circuit or as a part of a programmed microprocessor, depending on the frequency range of interest. *See* MICROPROCESSOR.

A digital intermediate-frequency stage is far more versatile and controllable than its analog counterpart. For example, while changing the selectivity of an analog filter may require switching to different physical signal paths, changing the selectivity of a digital filter requires only the change of stored numerical filter coefficients. Therefore, selective filtering, provided by a digital intermediate-frequency stage, can be more accurate than that of its analog counterpart, absolutely repeatable, and even program-controlled. Such filtering can be used to attenuate out-of-band noise and to sharply reduce in-band interference by use of individual (or a combination of) programmed, adaptive, and manually tuned notch filters. Furthermore, broadband noise can be greatly attenuated by using a digital adaptive noise suppressor. *See* AMPLIFIER.

Stanley A. White

Bibliography. R. C. Dorf (ed.), *The Electrical Engineering Handbook*, 2d ed., 1997; S. Gibilisco, *Handbook of Radio and Wireless Technology*, 1998; R. F. Graf, *Converter and Filter Circuits*, 1997; J. E. Paul, Adaptive Noise-Cancelling Receiver, U.S. Patent No. 4, 177, 430; M. S. Roden, *Analog and Digital Communications Systems*, 4th ed., 2000; M. E. Van Valkenburg (ed.), *Reference Data for Engineers: Radio, Electronics, Computer, and Communications*, 8th ed., 1996.

Intermediate vector boson

One of the three fundamental particles that transmit the weak force. (An example of a weak interaction process is nuclear beta decay.) These elementary particles—the W^+, W^-, and Z^0 particles—were discovered in 1983 in very high energy proton-antiproton collisions. It is through the exchange of *W* and *Z* bosons that two particles interact weakly, just as it is through the exchange of photons that two charged particles interact electromagnetically. The intermediate vector bosons were postulated to exist in the 1960s; however, their large masses prevented their production and study at accelerators until 1983. Their discovery was a key step toward unification of the weak and electromagnetic interactions. *See* ELECTROWEAK INTERACTION; ELEMENTARY PARTICLE; FUNDAMENTAL INTERACTIONS; WEAK NUCLEAR INTERACTIONS.

Production and detection. The *W* and *Z* particles are roughly 100 times the mass of a proton. Therefore, the experiment to search for the *W* and the *Z* demanded collisions of elementary particles at the highest available center-of-mass energy. Such very high center-of-mass energies capable of producing the massive *W* and *Z* particles were achieved with collisions of protons and antiprotons at the laboratory of the European Organization for Nuclear Research (CERN) near Geneva, Switzerland. The $p\bar{p}$ collisions were monitored in two underground experimental areas (UA1 and UA2). *See* PARTICLE ACCELERATOR; PARTICLE DETECTOR.

Properties of W and Z particles. Striking features of both the charged *W* and the Z^0 particles are their large masses. The charged boson (W^+ and W^-) mass is measured to be about 80 GeV/c^2, and the neutral boson (Z^0) mass is measured to be about 91 GeV/c^2. (For comparison, the proton has a mass of about 1 GeV/c^2.) Prior to the discovery of the *W* and the *Z*, particle theorists had met with some success in the unification of the weak and electromagnetic interactions. The electroweak theory as it is understood today is due largely to the work of S. Glashow, S. Weinberg, and A. Salam. Based on low-energy neutrino scattering data, which in this theory involves the exchange of virtual *W* and *Z* particles, theorists made predictions for the *W* and *Z* masses. The actual measured values are in agreement (within errors) with predictions. The discovery of the *W* and the *Z* particles at the predicted masses is an essential confirmation of the electroweak theory, one of the cornerstones of the standard model of particle physics. *See* STANDARD MODEL.

Only a few intermediate vector bosons are produced from 10^9 proton-antiproton collisions at a center-of-mass energy of 540 GeV. This small production probability per $p\bar{p}$ collision is understood to be due to the fact that the bosons are produced by a single quark-antiquark annihilation. The other production characteristics of the intermediate vector bosons, such as longitudinal and transverse momentum distributions (with respect to the $p\bar{p}$ colliding beam axis), provide support for this theoretical picture. *See* QUARKS.

Decay. The decay modes of the *W* and *Z* are well predicted. The simple decays $W^+ \rightarrow e^+\nu$, $W^- \rightarrow e^-\bar{\nu}$, $Z^0 \rightarrow e^+e^-$, and $Z^0 \rightarrow \mu^+\bar{\mu}$ are spectacular signatures of the intermediate vector bosons (**Fig. 1**). These leptonic decays are only a few percent of the total number of *W* and *Z* decays, however, because the *W* and *Z* also decay into

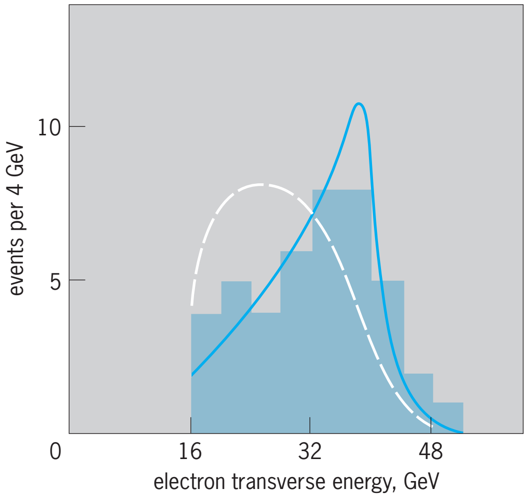


Fig. 1. Observation of the *W* particle in the UA1 data from the electron transverse energy distribution. The solid curve is the expected distribution for the decay, $W \rightarrow e\nu$, while the broken curve is the distribution expected for the decay of a different particle "X" into an electron and two neutrinos.

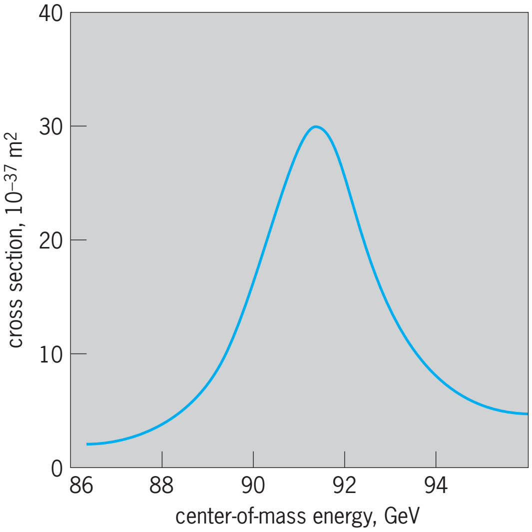


Fig. 2. Rates from the L3 experiment at LEP for the process $e^+e^- \rightarrow Z^0$ as a function of total center-of-mass energy of the electron plus positron. The rate is measured as a total cross section. (*After O. Adriani et al., Results from the L3 Experiment at LEP, Phys. Rep., 236:1–146, 1993*)

quark-antiquark pairs. The *W* particle also has an identifying feature in its decay, which is due to its production through quark-antiquark annihilation. The quark and antiquark carry intrinsic angular momentum (spin) that is conserved in the interaction. This means that the *W* is polarized at production; its spin has a definite orientation. This intrinsic angular momentum is conserved when the *W* decays, yielding a preferred direction for the decay electron or positron. This distribution is characteristic of a weak interaction process, and it determines the assignment of the spin quantum number of the *W* to be 1, as expected from theory. (Spin-1, or more generally integer-spin, particles are called bosons; another example is the photon.) *See* QUANTUM STATISTICS; SPIN (QUANTUM MECHANICS).

Results from electron-positron annihilation. In 1989, the Z^0 particle was produced in electron-positron annihilations at both the Stanford Linear Collider (SLC) and the CERN Large Electron-Positron (LEP) accelerator. The Z^0 was first observed in these experiments by its decay into a quark-antiquark ($q\bar{q}$) pair. The quark and antiquark are not observed as free particles; however, the energetic hadrons originating from them give a clear experimental signature. This process, $e^+e^- \rightarrow Z^0 \rightarrow q\bar{q}$, is the inverse of the fundamental process in which the Z^0 particle was discovered, $q\bar{q} \rightarrow Z^0 \rightarrow e^+e^-$, where the quark and antiquark are constituents of the energetic proton and antiproton. Electron-positron annihilations offer an exceptionally background-free method for the study of the Z^0 particle because it is created alone as a pure state. Thus, the LEP machine with its high collision rate is appropriately referred to as a Z^0 factory. Since the Z^0 particle couples to all known quarks and leptons, the study of its decays also provides information about the properties of all known particles.

The mass of the *Z* particle has been extremely accurately determined at LEP by varying the energy of the electrons and positrons by a few percent and observing the rate of the process $e^+e^- \rightarrow Z^0 \rightarrow q\bar{q}$ as a function of energy (**Fig. 2**). The result from four LEP experiments is that $m_Z = 91.1876$ GeV/c^2 with an experimental error of 0.0021 GeV/c^2. Many details of the decay of *Z* particles have been measured at LEP with high statistics. All are in agreement with the standard model. The final stages of LEP (LEP II) allowed an increase in machine energy so that pairs of W^+ and W^- were produced.

The *W* and *Z* particles have also been extensively studied at the Fermilab Tevatron proton-antiproton collider. Detection of the *W* particles was the key element in the observation of the top quark at Fermilab. The data from LEP II and Fermilab allowed accurate determination of the *W* mass. The present value is $m_W = 80.423$ GeV/c^2 with an experimental error of 0.039 GeV/c^2.

The Z^0 particle has an extremely short lifetime of about 10^{-25} s. This is due in part to the large number of channels into which it may decay. The result of this short lifetime is that the Z^0 particle is observed as a resonance having a natural width (Γ) of 2.495 GeV/c^2 (Fig. 2). There is a partial contribution to this width for each decay channel available to the Z^0. These partial widths can be calculated theoretically. The particles that are known to contribute to the Z^0 width are the electron, muon, and tau leptons ($Z^0 \rightarrow e^+e^-, \mu^+\mu^-$, or $\tau^+\tau^-$); the up, down, strange, charm, and bottom quarks ($Z^0 \rightarrow q\bar{q}$); and the electron-, muon-, and tau-neutrinos ($Z^0 \rightarrow \nu\bar{\nu}$). There is good agreement between the measured width and the sum of the theoretical partial widths for each known decay channel. The existence of possible new leptons, quarks, or neutrinos with masses less than about $m_Z/2$ is therefore ruled out. *See* LEPTON; NEUTRINO.

Now the focus in particle physics is to understand why the weak interaction mediators (*W* and *Z*) are massive while the electromagnetic interaction mediator (photon) is massless, since the two forces are unified. A new proton-proton accelerator, the Large Hadron Collider (LHC), will come into operation in 2008 to address this puzzle. The center-of-mass collision energy of 14 TeV and the high luminosity of 10^{34} cm^{-2} s^{-1} will provide, for the first time, interactions of *W* and *Z* particles at energies much larger than the mass energies of the bosons, providing the first window to probe the mechanism by which elementary particles obtain their masses.

James W. Rohlf

Bibliography. G. Arnison et al., Experimental observation of isolated large transverse energy electrons with associated missing energy at $\sqrt{s} = 540$ GeV, *Phys. Lett.*, 122B:103–116, 1983; G. Arnison et al., Experimental observation of lepton pairs of invariant mass around 95 GeV/c^2 at the CERN SPS collider, *Phys. Lett.*, 126B:398–410, 1983; The LEP collaborations: Electroweak parameters of the Z^0 resonance and the standard model, *Phys. Lett.*, B276:247–253, 1992; J. W. Rohlf, *Modern Physics from α to Z^0*, 1994.

Intermetallic compounds

All metal–metal compounds, including ordered, disordered, binary, and multicomponent. The metal–metal aspect of the definition is often relaxed to include some metal–metalloid compounds such as silicides and tellurides and certain compound semiconductors such as InSb. Such inclusion is appropriate since the phenomenology of many of these compounds is similar to the metal-metal ones. Intermetallic compounds constitute an intriguing field for scientific inquiry and a rewarding source of materials for diverse applications. *See* METAL; METALLOID; SEMICONDUCTOR.

Binary systems. Some of the formation characteristics of intermetallics may be appreciated from examination of the phase diagram such as that for the binary system Al-Ni in **Fig. 1**. Here it is seen that intermetallics can melt congruently (for example, AlNi), and they form by solid–liquid reactions (such as Al_3Ni, Al_3Ni_2, or $AlNi_3$) or interaction between solids (such as Al_3Ni_5). The phase diagram also shows that intermetallics can exhibit broad ranges of composition, extending on one or both sides of stoichiometry (for example, AlNi), or may show essentially no deviation from stoichiometry (for example, Al_3Ni). These same features also appear in ternary and higher-order systems. Stoichiometric deviation implies the presence of point defects (isolated atoms), including substitutions, vacancies, or interstitials. *See* CRYSTAL DEFECTS; EUTECTICS; NONSTOICHIOMETRIC COMPOUNDS; PHASE EQUILIBRIUM; SOLID-STATE CHEMISTRY; STOICHIOMETRY.

Crystal structure. Another characteristic of intermetallic compounds is that their crystal structures are idiosyncratic, not to be inferred from the structures of the component metals. Aluminium and nickel are face-centered-cubic (fcc) metals, but the five intermetallics shown in Fig. 1 have unique structures, none of them the simple fcc of the component metals. In this system, $AlNi_3$ exhibits an ordered variant of the fcc structure, a so-called superstructure (**Fig. 2***a*), but the others are quite different—orthorhombic, trigonal, or ordered body-centered-cubic (bcc). In this particular system, all the intermetallics remain ordered until the melting point is reached, but in other cases (for example, compounds in the Au-Cu system) the ordering disappears well before the melting point. In a few cases, a discrete compound exists in the system over a limited composition range and exhibits no ordering at all (for example, Ag_3Al). Some intermetallic crystal structures are quite simple with a low number of atoms per unit cell. For example, AlNi has the CsCl structure (ordered bcc) with only two atoms per unit cell (Fig. 2*b*), while others, such as $NaCd_2$ with a simple formula, have very complex structures with hundreds of atoms per unit cell. In general, intermetallic crystal structures tend to be high-symmetry, high-space-filling, and high-coordination-number. Although some 2750 intermetallic crystal structures are known, ∼5000 are potentially possible. The 100 most common crystal structures account for about 72% of all known intermetallics (∼25,000). The particular structure adopted by a pair of metal atoms depends on their relative size, the atomic number factor, the electrochemical factor, the valence electron factor, and the angular valence–orbital factor. The relative importance of these factors varies from case to case. *See* ATOM; COORDINATION NUMBER; CRYSTAL STRUCTURE; VALENCE.

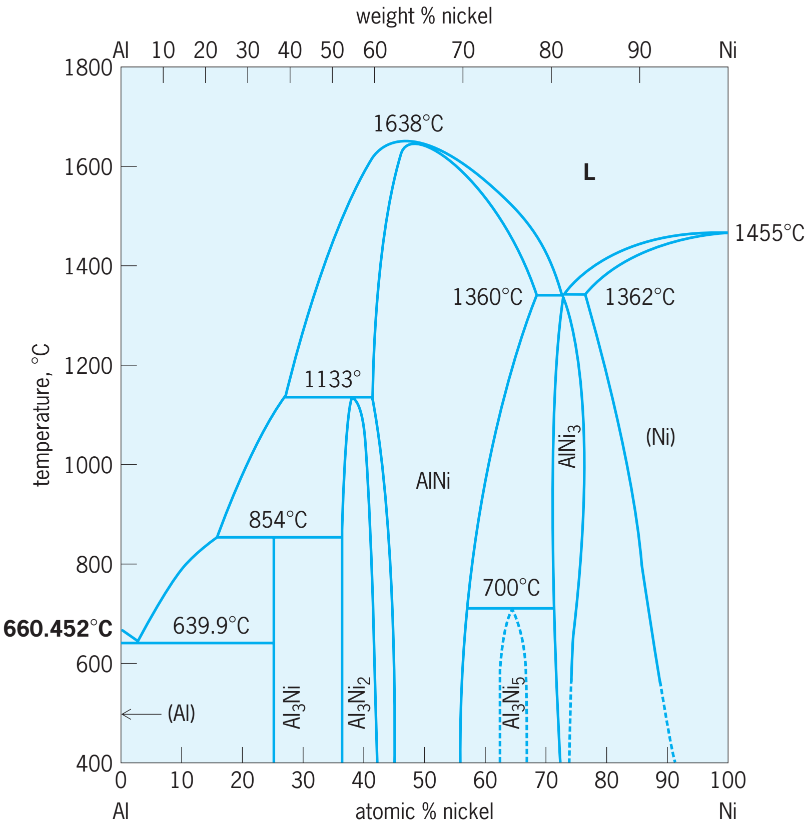


Fig. 1. Al-Ni phase diagram. Note that the peritectic formation of $AlNi_3$ lies only 2° above the eutectic between AlNi and $AlNi_3$. (*After J. H. Westbrook, ed., Handbook of Binary Phase Diagrams, Genium Publishing, 2005*)

Ternary systems. For ternary systems, the situation is even more complex and little is known. About 15,000 ternary compounds have been identified, but this is only about 3% of the expected number of the possible ternary intermetallics. Ternary compounds are not just of conceptual interest but also of commercial significance, as seen by the important magnetic phase, $Nd_2Fe_{14}B$, and the embrittling G-phase, $Ni_{16}Ti_6Si_7$, sometimes encountered in high-temperature alloys. Complications do not end with ternaries. A seven-component intermetallic ($Cu_{11.5}Zn_{23}Ga_{7.5}Ge_8As_{16}Se_{30}Br_4$) with a disordered zinc blende structure has been successfully synthesized. Quasicrystals, structures lacking long-range order but exhibiting quasiperiodicity and icosahedral or decagonal structures, and intermetallic clusters (for example, $Ba_{16}Na_{204}Sn_{310}$) are also intriguing intermetallics. *See* ALLOY; ALLOY STRUCTURES; QUASICRYSTAL.

Formation. Intermetallic compounds can be prepared in various ways, depending on the particular compound and the intended product. Possible methods include melting and casting, powder metallurgy,

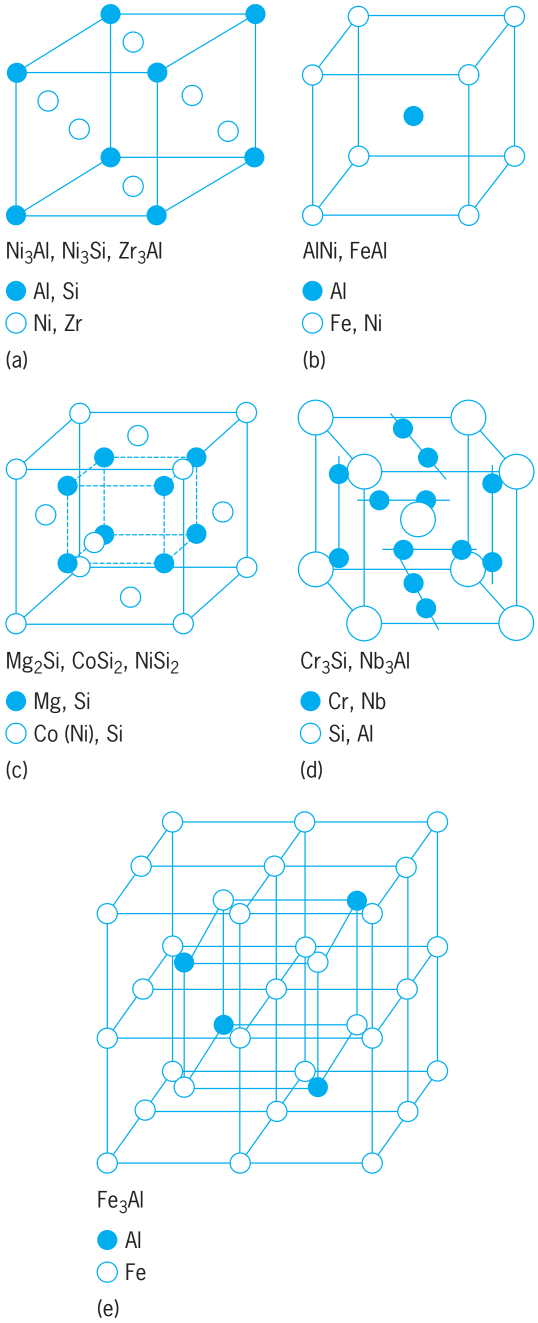


Fig. 2. General cubic lattices of aluminides and silicides. (*a*) $L1_2$ [cP4], (*b*) B2 [cP2], (*c*) C1 [cF12], (*d*) A15 [cP8], (*e*) $D0_3$ [cF16]. (*After K. H. J. Buschow et al., eds., The Encyclopedia of Materials: Science and Technology, vol. 5, p. 4177, Pergamon, 2001*)

mechanical alloying, diffusion, nanotech synthesis, and molecular-beam epitaxy. They also can occur naturally as minerals—for example, braithauptite (NiSb), domeykite (Cu_3As), froodite ($PdBi_2$), drysdallite ($MoSe_2$), nigglite (PtSn), petzite [(Ag,Au)Te_2], kolymite (Cu_7Hg_6), and hapkeite (Fe_2Si), to cite a single example in each family. *See* DIFFUSION; MECHANICAL ALLOYING; POWDER METALLURGY.

Applications. Practical use of intermetallics goes far back in time. In prehistoric times, meteoritic NiFe was used for tools and Cu_3Au (tumbaga) for fishhooks. In Roman times, CuSn was used for mirrors and Cu_3Sn in bronze swords and bells. In the eighteenth and nineteenth centuries, CuZn was used for ship sheathing, SbSn in printing type and bearings, and Ag_2Hg_3 and Sn_8Hg in dental amalgams. Today, intermetallics are used in many high-tech applications, such as Ni_3Al in jet engine alloys, $FeZn_3$ in automotive galvanized steel sheet, FeAl as a resistive heating element, GaAs in light-emitting diodes and semiconducting substrates, $LaNi_5$ as a battery electrode, (Tb,Dy)Fe_2 as a magnetoelastic transducer, and $Ni_{0.46}Ti_{0.54}$ as shape memory alloy in the biomedical field. Other specialized applications include protective coatings, catalysts, heat storage materials, thermoelectrics, pyrophoric alloys, and jewelry. The market for intermetallics in the United States has been estimated at \$10 billion per year.

Other intermetallics. There are a few cases where intermetallics exhibit truly exceptional behavior. $LaGa_3$ is a "plastic" crystal with orientational ordering. Cs_4Pb_4 undergoes a two-step melting (first loss of rotational order, then loss of positional order). CeSb has the most complex magnetic system known (>17 magnetic structures). Colored intermetallics include CoGa (yellow), $CoSi_2$ (dark blue), $AuAl_2$ (violet), $AuLi_2Sn$ (pink), and FeAl (brown), to name some. Pd_8U is a thermoelectric with the very large positive thermopower for a metal. $SmMn_2Ge_2$ is a magnetoresistor. NaTl is a substance having ionic, covalent, and metallic bonds, all in the same crystal. Mg_3Bi_2 is a superionic conductor with mobile Mg^{2+} cations in a metallic framework. And Cu_2MnAl is a ferromagnet with no ferromagnetic constituents.

J. H. Westbrook

Bibliography. F. R. de Boer and D. G. Pettifor (eds.), *Cohesion and Structure*, North-Holland, 1991; M. Doyama and M. Yabe, *Databook of Intermetallic Compounds*, Science Forum, Tokyo, 1988; R. D. Dudley and P. D. Desai, *Properties of Intermetallic Alloys III*, CINDAS, Purdue University, 1995; R. L. Fleischer and J. H. Westbrook, *Intermetallic Compounds; Principles and Practice*, Wiley, vol. 1, 1996, vol. 2, 1996, vol. 3, 2002; B. P. Gilp and P. D. Desai, *Properties of Intermetallic Alloys II*, CINDAS, Purdue University, 1994; C. T. Liu, R. W. Cahn, and G. Sauthoff (eds.), *Ordered Intermetallics: Physical Metallurgy and Mechanical Behavior*, Kluwer, 1992; N. M. Matveeva and E. V. Kozlov, *Ordered Phases in Metallic Systems*, Nova Science, 1996; National Materials Advisory Board, *Intermetallic Alloy Development: A Program Evaluation*, NMAB-487-1, National Academy Press, 1997; J. E. Payne and P. D. Desai, *Properties of Intermetallic Alloys I*, CINDAS, Purdue University, 1994; H. Ringpfeil (ed.), *Intermetallische Phasen*, VEB Deutscher Verlag, Leipzig, 1976; G. Sauthoff, *Intermetallics*, VCH, 1995; N. S. Stoloffand and V. K. Sikka, *Physical Metallurgy and Processing of Intermetallic Compounds*, Chapman and Hall, 1996; P. Villars, *Pearson's Handbook: Crystallographic Data for Intermetallic Phases*, Desk ed., 2 vols., ASM, 1997; H. Warlimont (ed.), *Order-Disorder Transformations in Alloys*, Springer, 1974.

Intermolecular forces

Attractive or repulsive interactions that occur between all atoms and molecules. Intermolecular forces become significant at molecular separations of about 1 nanometer or less, but are much weaker than the forces associated with chemical bonding. They are important, however, because they are responsible for many of the physical properties of solids, liquids, and gases. These forces are also largely responsible for the three-dimensional arrangements of biological molecules and polymers.

Attractive and repulsive forces. The observation that gases condense into liquids and solids at low temperatures, rather than continuing to occupy all the space available to them, is evidence of a force of attraction between their molecules. This attractive force is overcome only when the thermal energy is high enough. The further observation that liquids and solids occupy a finite volume and cannot be compressed easily is evidence of a repulsive force that comes into play when the molecules are close together.

Accordingly, we can deduce that the energy of interaction between two molecules or atoms takes a form like that in **Fig. 1**, which shows the interaction between argon atoms as deduced from experimental data. Argon atoms are electrically neutral and do not form chemical bonds with each other. When two argon atoms are far apart, the energy of interaction is essentially zero, and the atoms do not exert a force on each other. (The force is the negative gradient, or slope, of the potential energy.) As the two atoms approach, the energy between them decreases, reaching a minimum at a distance of 0.38 nm and then increasing again. At distances greater than 0.38 nm, the energy increases with increasing distance, and the force between the molecules is attractive; at shorter distances, the energy increases rapidly with decreasing distance, and the force becomes strongly repulsive. At the minimum, the interaction energy between the argon atoms is about 1.2 kJ/mol, which may be compared with typical chemical bond energies, of the order of 300 kJ/mol. *See* ARGON; CHEMICAL BONDING.

Description. Intermolecular forces can be classified into several types, of which two are universal (as illustrated by the same example of two interacting argon atoms). The attractive force known as dispersion arises from the quantum-mechanical fluctuation of the electron density around the nucleus of each atom. At long distances—greater than 1 nm or so—the electrons of each atom move independently of the other, and the charge distribution is spherically symmetric (**Fig. 2***a*). At shorter distances, an instantaneous fluctuation of the charge density in one atom can affect the other. If the electrons of one atom move briefly to the side nearer the other (Fig. 2*b*), the electrons of the other atom are repelled to the far side. In this configuration, both atoms have a small dipole moment, and they attract each other electrostatically. At another moment, the electrons may move the other way, but their motions are correlated so that an attractive force is maintained on average. Molecular orbital theory shows that the electrons of each atom are slightly more likely to be on the side nearer to the other atom, so that each atomic nucleus is attracted by its own electrons in the direction of the other atom. Fritz London demonstrated in 1930 that at long range the dispersion energy is

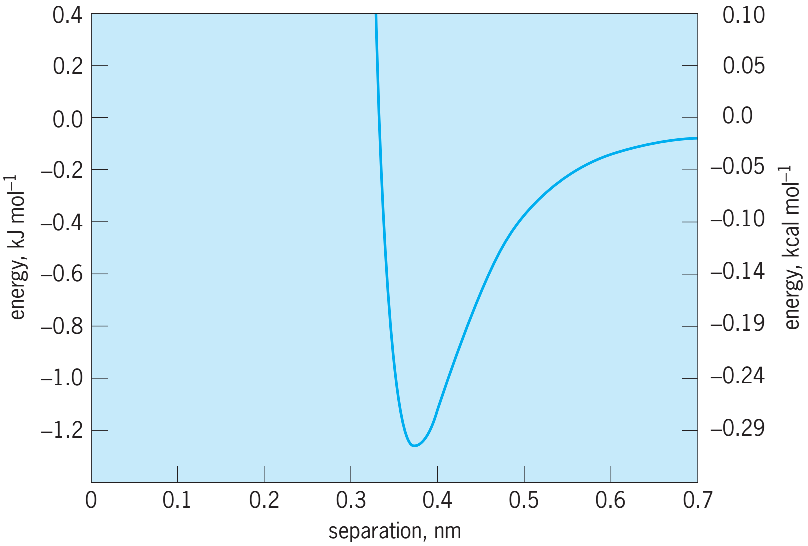


Fig. 1. Intermolecular potential energy of two argon atoms. (*After R. A. Aziz and H. H. Chen, An accurate intermolecular potential for Ar, J. Chem. Phys., 67:5719-5726, 1977*)

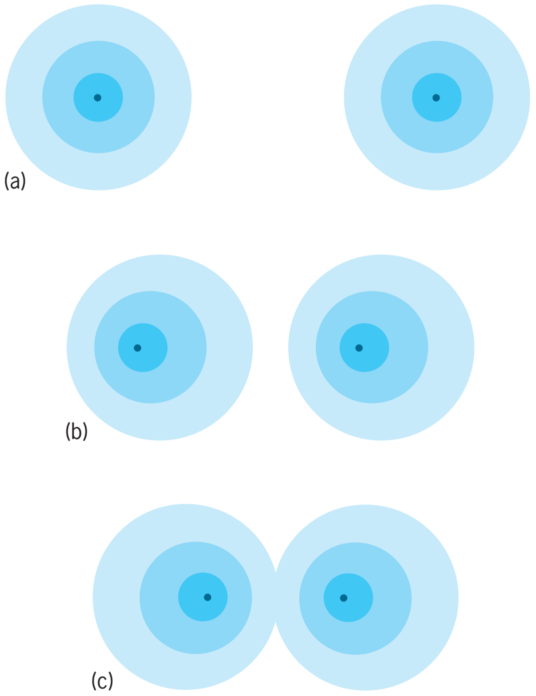


Fig. 2. Schematic diagram of intermolecular interaction. (*a*) There is negligible interaction between atoms that are 1 nm apart. (*b*) For atoms separated by about 0.8 nm or less, attractive dispersion forces result from correlated fluctuations of the electron charge distributions of the atoms. (*c*) For atoms closer than 0.3 nm or so, exchange effects cause a distortion of the charge distribution that leads to repulsion.

proportional to the inverse sixth power of the distance between the atoms. This London dispersion energy is larger for atoms and molecules whose electron densities fluctuate freely—in other words, those atoms and molecules that are easily polarized by external electric fields. *See* DIPOLE MOMENT; MOLECULAR ORBITAL THEORY.

At small separations the electron clouds can overlap, and repulsive forces arise. These forces are described as exchange–repulsion and are a consequence of the Pauli exclusion principle, a quantum-mechanical effect which prevents electrons from occupying the same region of space simultaneously (Fig. 2*c*). To accommodate it, electrons are squeezed out from the region between the nuclei, which repel each other as a result. Each element can be assigned, approximately, a characteristic van der Waals radius; that is, when atoms in different molecules approach more closely than the sum of their radii, the repulsion energy increases sharply. It is this effect that gives molecules their characteristic shape, leading to steric effects in chemical reactions. *See* EXCLUSION PRINCIPLE; STERIC EFFECT (CHEMISTRY).

The other important source of intermolecular forces is the electrostatic interaction. When molecules are formed from atoms, electrons flow from electropositive atoms to electronegative ones, so that the atoms become somewhat positively or negatively charged. In addition, the charge distribution of each atom may be distorted by the process of bond formation, leading to atomic dipole and quadrupole moments. The electrostatic interaction between these is an important source of intermolecular forces, especially in polar molecules, but also in molecules that are not normally thought of as highly polar. The interaction between two ethyne (acetylene) molecules, for example, is dominated by the quadrupole moment associated with the carbon-carbon triple bond. A further effect is that the electrostatic field of a molecule may cause polarization of its neighbors, and this leads to a further induction contribution to the intermolecular interaction. An important feature is that an induction interaction can often polarize both molecules in such a way as to favor interactions with further molecules, leading to a cooperative network of intermolecular attractions. This effect is important in the network structure of water and ice. *See* WATER.

An example of intermolecular forces is the hydrogen bond, in which a hydrogen atom attached to an electronegative atom such as oxygen or nitrogen is weakly bonded to another electronegative atom in the same molecule or a different molecule. The dominant attractive contribution in a hydrogen bond is the electrostatic interaction, but there are also significant contributions from dispersion and induction. *See* HYDROGEN BOND.

Occurrence. Intermolecular forces are responsible for many of the bulk properties of matter in all its phases. A realistic description of the relationship between pressure, volume, and temperature of a gas must include the effects of attractive and repulsive forces between molecules. The viscosity, diffusion, and surface tension of liquids are examples of physical properties which depend strongly on intermolecular forces. Intermolecular forces are also responsible for the ordered arrangement of molecules in solids, and account for their elasticity and properties (such as the velocity of sound in materials). The detailed description of the effects of intermolecular forces in condensed phases is complicated, since every molecule interacts not just with its immediate neighbors but with more distant ones too. *See* DIFFUSION; ELASTICITY; GAS; SURFACE TENSION; VAN DER WAALS EQUATION; VIRIAL EQUATION; VISCOSITY.

Intermolecular forces are important between atoms within a molecule. The repulsions between atoms constrain the possible three-dimensional structures that can be taken up by a large molecule such as a polymer or protein. Hydrogen bonds are especially important in biological molecules. The energy associated with a hydrogen bond is typically of the order of 10–20 kJ/mol, so such bonds are much weaker than normal chemical bonds. In a protein, however, there are many hydrogen bonds between the amino acid residues, and they are responsible for the stability of the alpha helix and other elements of protein structure. The twin strands of DNA are composed of adenine, cytosine, guanine, and thymine molecules, connected by a phosphate–sugar backbone. An adenine molecule in one strand can pair only with a thymine in the other, and cytosine can pair only with guanine. This selectivity, which is responsible for the duplication of the genetic code, arises from the pattern of hydrogen bonds. *See* DEOXYRIBONUCLEIC ACID (DNA); PROTEIN.

Atoms and molecules may be held to a solid surface by intermolecular forces. This weak bonding, called physisorption, has many important applications. The trapping of molecules from the gas phase onto cooled surfaces is the basis of pumps for producing high vacuums. Undesirable odors or colors in food or water may sometimes be removed by filters which capture the offending contamination by physisorption. The selective adsorption of molecules by surfaces is a useful method for separation of mixtures of molecules. Some solid catalysts function by attracting molecules from the gas phase onto the surface, where they can react with each other. *See* ADSORPTION; CATALYSIS; CHROMATOGRAPHY; VACUUM PUMP.

Experimental techniques. Because of their importance, intermolecular forces have been studied extensively for many years. Until the early 1970s, most of the information on intermolecular forces was inferred from the study of matter in bulk. For example, measurements of the viscosity of gases or the crystal structure of solids were used for this purpose. However, such properties depend rather indirectly on the intermolecular forces, and the conclusions of such studies were often ambiguous. However, the intermolecular potential energy curve shown in Fig. 1 was derived from experimental data on the viscosity, thermal conductivity, diffusion, and second pressure virial coefficient of argon gas, together with spectroscopic data on the argon dimer, Ar_2.

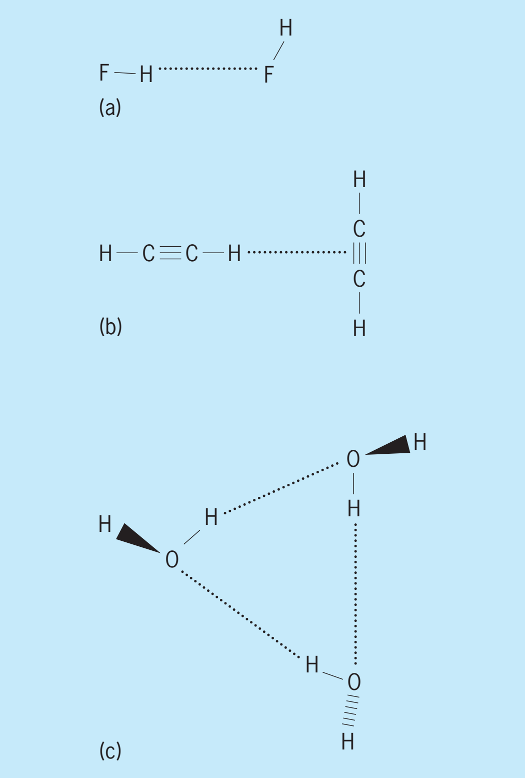


Fig. 3. Structures of some van der Waals molecules. (*a*) Hydrogen fluoride dimer. (*b*) Ethyne dimer. (*c*) Water trimer; two of the non-hydrogen-bonded H atoms are on one side of the plane of the O atoms, and the third is on the other.

Modern methods study interactions between individual molecules. In molecular-beam collision experiments, low-density streams of atoms or molecules are directed so that individual particles collide. The way in which the molecules rebound from the collision depends on their initial velocities, which can be controlled, and on the forces acting between them. Information about intermolecular forces can be extracted from the experimental data. *See* MOLECULAR BEAMS.

A method which has been increasingly important in recent years, with the development of sophisticated spectroscopic techniques using lasers, involves van der Waals molecules. A van der Waals molecule is a cluster of two or more atoms or molecules in which the molecules retain their individual identity and are held together by intermolecular forces. Such molecules are usually made by expanding a mixture of gases, usually with an inert carrier gas such as argon, into a high vacuum through a narrow nozzle. The resulting molecular beam contains van der Waals molecules, which can be studied spectroscopically. Very detailed information about intermolecular forces can be obtained in this way. In the case of the Ar-HCl cluster, for instance, the interaction energy is now known accurately and in detail as a function of the separation and orientation of the argon and hydrogen chloride components. Other van der Waals molecules have been studied in this way, including clusters of up to six water molecules, $(HF)_2$, $(C_2H_2)_2$, and $(C_2H_2)_3$. The structures of some van der Waals molecules are shown in **Fig. 3**. The T-shaped structure of ethyne dimer is a consequence of the quadrupole–quadrupole interaction mentioned above, while the bent structure of HF dimer is also attributed to electrostatic forces. Indeed, it has become apparent in recent decades that the electrostatic forces are dominant in determining the structures of van der Waals molecules, primarily because they are repulsive in some orientations and attractive in others, and so are much more orientation-dependent than other intermolecular forces. *See* LASER.

Many of these molecules can rearrange into equivalent but different structures. The HF dimer, for instance, can rearrange in the way shown in **Fig. 4**, the hydrogen atom in the hydrogen bond moving aside and the other taking its place. The symmetrical intermediate structure is only about 4 kJ/mol higher in energy than the initial and final structures, and tunneling can occur between them, leading to splittings in the spectrum that can be measured. These tunneling splittings provide further information about the nature of the intermolecular potential energy surface. *See* MOLECULAR STRUCTURE AND SPECTRA.

Computational methods. An important source of information about intermolecular potential energy surfaces is quantum theory. The underlying general theory of intermolecular forces has been understood since the 1930s, but only in recent years has it become possible to explore it quantitatively and to understand its finer details. Calculations ab initio, starting from Schrödinger's equation without any empirical information and using quantum-mechanical perturbation theory, can provide accurate details of all features of the forces between molecules, though this approach is limited to relatively small molecules. In this way, it is possible to obtain a deeper

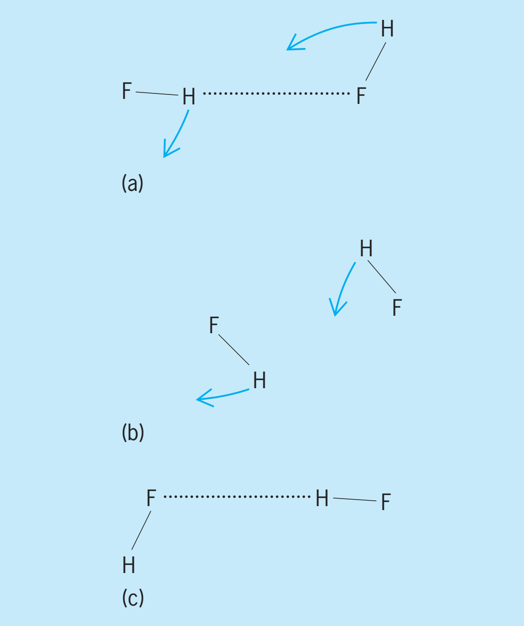


Fig. 4. Donor–acceptor exchange in HF dimer.

understanding of the quantum-mechanical origin of intermolecular forces and to calculate them to an accuracy comparable with experiment. *See* NONRELATIVISTIC QUANTUM THEORY; QUANTUM CHEMISTRY.

Anthony J. Stone

Bibliography. J. Israelachvili, *Intermolecular and Surface Forces*, Academic Press, 1992; J. S. Rowlinson, *Cohesion: A Scientific History of Intermolecular Forces*, Cambridge University Press, 2002; A. J. Stone, *The Theory of Intermolecular Forces*, Oxford University Press, 1996; D. J. Wales, *Energy Landscapes*, Cambridge University Press, 2003.

Internal combustion engine

A prime mover that burns fuel inside the engine, in contrast to an external combustion engine, such as a steam engine, which burns fuel in a separate furnace. *See* ENGINE.

Most internal combustion engines are spark-ignition gasoline-fueled piston engines. These are used in automobiles, light- and medium-duty trucks, motorcycles, motorboats, lawn and garden equipment, and light industrial and portable power applications. Diesel engines are used in automobiles, trucks, buses, tractors, earthmoving equipment, as well as marine, power-generating, and heavier industrial and stationary applications. This article describes these types of engines. For other types of internal combustion engines *see* GAS TURBINE; ROCKET PROPULSION; TURBINE PROPULSION.

The aircraft piston engine is fundamentally the same as that used in automobiles but is engineered for light weight and is usually air-cooled. *See* RECIPROCATING AIRCRAFT ENGINE.

Engine types. Characteristics common to all commercially successful internal combustion engines include (1) the compression of air, (2) the raising of air temperature by the combustion of fuel in this air at its elevated pressure, (3) the extraction of work from the heated air by expansion to the initial pressure, and (4) exhaust.

Four-stroke cycle. William Barnett first drew attention to the theoretical advantages of combustion under compression in 1838. In 1862 Beau de Rochas published a treatise that emphasized the value of combustion under pressure and a high ratio of expansion for fuel economy; he proposed the four-stroke engine cycle as a means of accomplishing these conditions in a piston engine (**Fig. 1**). The engine requires two revolutions of the crankshaft to complete one combustion cycle. The first engine to use this cycle successfully was built in 1876 by N. A. Otto. *See* OTTO CYCLE.

Otto's engine, like almost all internal combustion engines developed at that period, burned coal gas mixed in combustible proportions with air prior to being drawn into the cylinder. The engine load was generally controlled by throttling the quantity of charge taken into the cylinder. Ignition was by a device such as an external flame or an electric spark, so that the timing was controllable. These are essential features of what has become known as the Otto or spark-ignition combustion cycle.

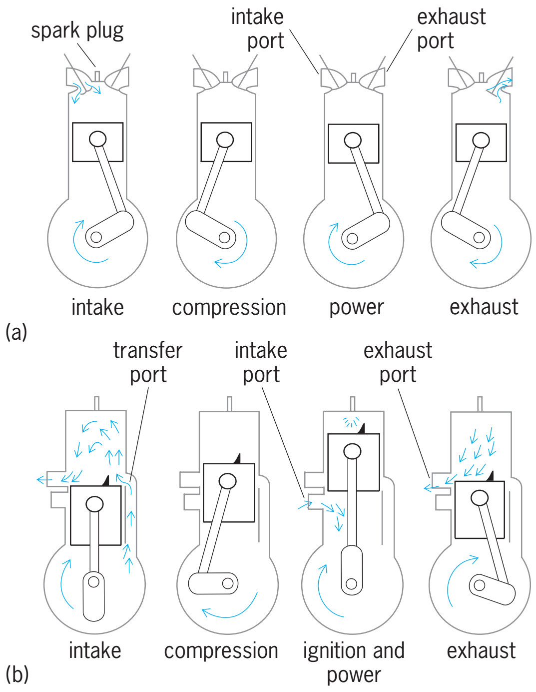


Fig. 1. Engine cycles (*a*) The four strokes of a four-stroke engine cycle. On intake stroke, the intake valve (left) has opened and the piston is moving downward, drawing air and gasoline vapor into the cylinder. On compression stroke, the intake valve has closed and the piston is moving upward, compressing the mixture. On power stroke, the ignition system produces a spark that ignites the mixture. As it burns, high pressure is created, which pushes the piston downward. On exhaust stroke, the exhaust valve (right) has opened and the piston is moving upward, forcing the burned gases from the cylinder. (*b*) Three-port two-cycle engine. The same action is accomplished without separate valves and in a single rotation of the crankshaft.

Two-stroke cycle. In 1878 Dougald Clerk developed the two-stroke engine cycle by which a similar combustion cycle required only one revolution of the crankshaft. In this cycle, exhaust ports in the cylinder were uncovered by the piston as it approached the end of its power stroke. A second cylinder then pumped a charge of air to the working cylinder through a check valve when the pump pressure exceeded that in the working cylinder.

In 1891 Joseph Day simplified the two-stroke engine cycle by using the crankcase to pump the required air. The compression stroke of the working piston draws the fresh combustible charge through a check valve into the crankcase, and the next power stroke of the piston compresses this charge. The piston uncovers the exhaust ports near the end of the power stroke and slightly later uncovers intake ports opposite them to admit the compressed charge from the crankcase. A baffle is usually provided on the piston head of small engines to deflect the charge up one side of the cylinder to scavenge the remaining burned gases down the other side and out the exhaust ports with as little mixing as possible.

Modern engines using this two-stroke cycle have a third cylinder port known as the transfer port (Fig. 1*b*), instead of the crankcase check valve used by Day. Small engines of this type are widely used where fuel economy is not as important as mechanical simplicity and light weight. They do not need mechanically operated valves, and they develop one power impulse per cylinder for each crankshaft revolution.

Two-stroke-cycle engines do not develop twice the power of four-stroke-cycle engines with the same size of working cylinders at the same number of revolutions per minute (rpm). Principal reasons are (1) reduction in effective cylinder volume due to the piston movement required to cover exhaust ports; (2) appreciable mixing of burned (exhaust) gases with the combustible mixture; and (3) loss of some combustible mixture through the exhaust ports with the exhaust gases.

Otto-cycle engines. In the idealized four-stroke Otto cycle, combustion is instantaneous and at constant volume. This simplifies thermodynamic analysis, but combustion takes time. Gas pressure during the four strokes of the Otto cycle varies with the piston position as shown by the typical indicator card in **Fig. 2***a*. This is a pressure–volume (PV) card for an 8.7:1 compression ratio.

Engine power. To simplify calculations of engine power, the average net pressure during the working stroke, known as the mean effective pressure (mep), is frequently used. It may be obtained from the average net height of the card, which is found by measurement of the area and then division of this area by its length. Similar pressure–volume data may be plotted on logarithmic coordinates as in Fig. 2*b*, which develops expansion and compression relations as approximately straight lines. The slopes show the values of exponent n to use in equations for PV relationships.

The rounding of the plots at peak pressure, with the peak developing after the piston has started its power stroke, even with the spark occurring before the piston reaches the end of the compression stroke, is due to the time required for combustion. Changes in design can vary charge turbulence in the compression space prior to and during combustion. The greater the turbulence, the faster the combustion and the lower the antiknock or octane number required of the fuel, or the higher the compression ratio that may be used with a given fuel without knocking. The amount to which the turbulence can be raised is limited by the increased rate of pressure rise, which increases engine roughness. This must not exceed a level acceptable for automobile or other service. *See* AUTOMOBILE; AUTOMOTIVE ENGINE; COMBUSTION CHAMBER; COMPRESSION RATIO; MEAN EFFECTIVE PRESSURE; OCTANE NUMBER; SPARK KNOCK.

Detonation of a small part of the charge in the cylinder, after most of the charge has burned progressively, causes knock. This limits the compression ratio of an engine with a given fuel.

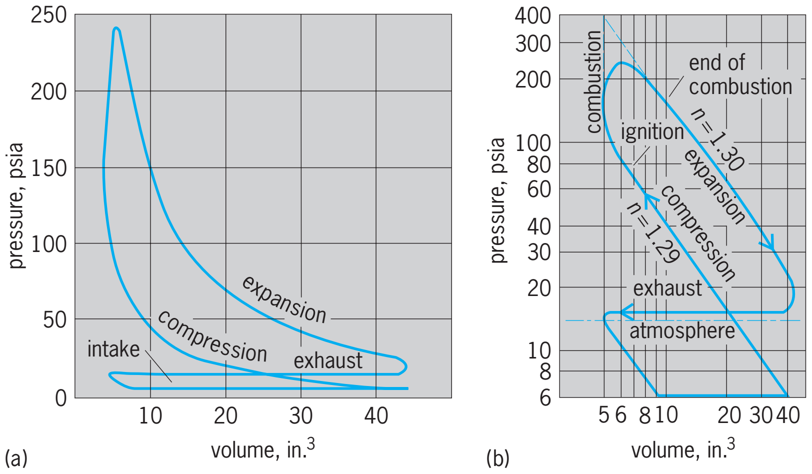


Fig. 2. Typical Otto-cycle pressure–volume indicator card plotted on (*a*) rectangular coordinates and (*b*) logarithmic coordinates.

Compression ratio. According to classical thermodynamic theory, thermal efficiency η of the Otto combustion cycle is given by Eq. (1), where the com-

$$\eta = 1 - \frac{1}{r^{n-1}} \qquad (1)$$

pression ratio r_c and expansion ratio r_e are the same ($r_c = r_e = r$). When theory assumes atmospheric air in the cylinder for extreme simplicity, exponent n is 1.4. Efficiencies calculated on this basis are almost twice as high as measured efficiencies. Logarithmic diagrams from experimental data show that n is about 1.3 (Fig. 2*b*). Even with this value, efficiencies achieved in practice are less than given by Eq. (1), because it assumes instantaneous combustion and 100% volumetric efficiency. This exponent should vary with the fuel-air mixture ratio, and to some extent with the compression ratio. For an 8:1 compression ratio, the exponent should vary from about 1.28 for a stoichiometric (chemically correct) mixture to about 1.31 for a lean mixture. Actual practice gives even lower thermal efficiencies. This is because of the assumed instantaneous changes in cyclic pressure (during combustion and exhaust) and the disregard of heat losses to the cylinder walls.

A change in compression ratio causes little change in the mechanical efficiency, or the volumetric efficiency resulting from raising the compression ratio provides a corresponding increase in torque or mean effective pressure. This is frequently of more practical importance than the actual efficiency increase. *See* THERMODYNAMIC CYCLE.

Engine load has little effect on indicated thermal efficiency, provided the fuel–air ratio remains constant and the ignition time is suitably advanced at reduced loads. This compensates for the slower rate of burning that results from dilution of the combustible charge with the larger percentages of burned gases remaining in the combustion space and the reduced turbulence at lower speeds.

High compression improves fuel economy because of improved thermal efficiency. However, the increased peak combustion temperature increases

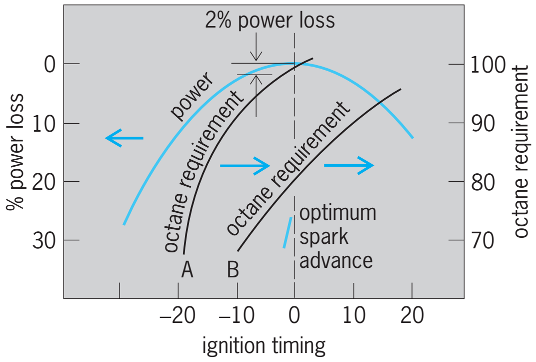


Fig. 3. Effects of advancing or retarding ignition timing from optimum on engine power and resulting octane requirement of fuel in an experimental engine with a combustion chamber having typical turbulence (A) and a highly turbulent design (B) with the same compression ratio. Retarding the spark 7° for 2% power loss reduced octane requirement from 98 to 93 for design A.

emissions of oxides of nitrogen in the exhaust gas. *See* AIR POLLUTION; SMOG.

Ignition timing. High thermal efficiency is obtained from high compression ratios at part loads, where engines normally run at automobile cruising speeds, with optimum spark advance. To avoid knock on available gasolines at wide-open throttle, a reduced or compromise spark advance is used. The tendency of an engine to knock at wide-open throttle is reduced appreciably when the spark timing is reduced 5–10° from optimum (**Fig. 3**). Advancing or retarding the spark timing from optimum results in an increasing loss in mean effective pressure for any normal engine, as shown by the heavy curve in Fig. 3. The octane requirement falls rapidly as the spark timing is retarded, the actual rate depending on the nature of the gasoline as well as on the combustion chamber design. Curves A and B show the effects on a given gasoline of the use of moderate- and high-turbulence combustion chambers, respectively, with the same compression ratio. Because the curve for mean effective pressure is relatively flat near optimum spark advance, retarding the spark for a 1–2% loss is normally acceptable because of the reduction in octane requirement.

In addition to the advantages of the higher compression ratio at cruising loads with optimum spark advance, the compromise spark at full load may be advanced toward optimum with higher-octane fuels. Then there is a corresponding increase in full-throttle mean effective pressure.

Many automotive engines have an electronic engine control system with an onboard microprocessor that controls spark timing electronically. The microprocessor continuously adjusts ignition timing for optimum fuel economy and drivability, while minimizing exhaust emissions. The microprocessor may be capable of retarding the timing if spark knock occurs. This allows the benefits of using a higher compression ratio and gasoline with a lower octane number without danger of engine-damaging detonation. In some systems, a digital map stored in memory provides a wide range of predetermined ignition settings (**Fig. 4**). *See* CONTROL SYSTEMS; IGNITION SYSTEM; MICROPROCESSOR.

Fuel–air ratio. A fuel-air mixture richer than that which develops maximum knock-free mep will permit use of higher compression ratios. However, the benefits derived from compromise or rich mixtures vary so much with mixture temperature and the sensitivity of the octane value of the particular fuel to temperature that this method is not generally practical. Nevertheless, piston-type aircraft engines may use fuel–air mixture ratios of 0.11 or even higher during takeoff, instead of about 0.08, which normally develops maximum mep in the absence of knock.

In automotive engines with an electronic engine control system, the microprocessor usually controls the amount of fuel delivered by either a feedback carburetor or a single-point (throttle-body) or a multipoint (port) electronic fuel-injection system. This maintains the mixture at or near the stoichiometric ratio, which minimizes exhaust emissions of hydrocarbons, carbon monoxide, and oxides of nitrogen. However, spark-ignition engines deliver maximum power with an air deficiency of 0–10%, and minimum fuel consumption with about 10% excess air (**Fig. 5**).

Stroke–bore ratio. The ratio of the length of the piston stroke to the diameter of the cylinder bore has no appreciable effect on fuel economy and friction at corresponding piston speeds. Practical advantages that result from the short stroke include the greater rigidity of crankshaft from the shorter crank cheeks, with crankpins sometimes overlapping main bearings, and the narrower as well as lighter cylinder block that is possible. However, the higher rates of crankshaft rotation for an equivalent piston speed necessitate greater valve forces and require stronger valve springs. Also, the smaller depth of the compression space for a given compression ratio increases the surface-to-volume ratio and the proportion of heat lost by radiation during combustion. In automotive engines, stroke–bore ratios have decreased over the years.

Valve timing. The times of opening and closing the valves of an engine in relation to piston position are usually selected to develop maximum power over a desired speed range at wide-open throttle. The

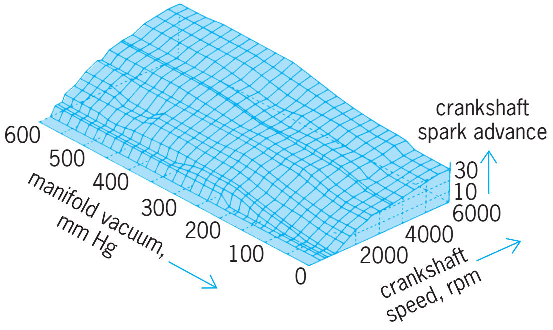


Fig. 4. Three-dimensional ignition map showing 576 timing points stored in the memory of a microprocessor system for a four-cylinder automotive engine. 1 mmHg = 133 Pa. (*Ford Motor Co.*)

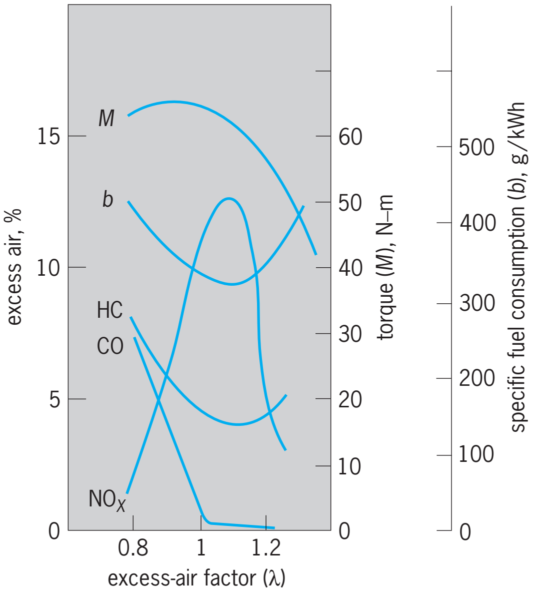


Fig. 5. Graph showing how the excess-air factor affects exhaust-gas [composition carbon monoxide (CO), nitrogen oxides (NO_x), and hydrocarbons (HC)], torque (*M*), and specific fuel consumption (*b*) in the part-load range of an automotive spark-ignition engine running at a constant midrange speed and cylinder charge. (*Robert Bosch Corp.*)

timing of these events is usually expressed as the number of degrees of crankshaft rotation before or after the piston reaches the end of one of its strokes.

Because of the time required for the burned gas to flow through the exhaust valve at the end of the power stroke of a piston, the exhaust valve usually starts opening considerably before the end of the stroke. If the valve opens when the piston is nearer the lower end of its stroke, power is lost at high engine speeds because the piston on its exhaust stroke has to move against gas pressure remaining in the cylinder. If the valve opens before necessary, the burned gas is released while it is still at sufficient pressure to increase the work done on the piston.

For any engine, there is an optimum time for opening the exhaust valve that will develop the maximum power at some particular speed. The power loss at other speeds does not increase rapidly. Therefore when an engine is throttled at part load, there is less gas to discharge through the exhaust valve and less need for the valve to be opened as early as at wide-open throttle.

The timing of intake valve events is normally selected to trap the largest possible quantity of combustible mixture (air in a diesel engine) in the cylinder when the valve closes at some desired engine speed and at wide-open throttle. The intermittent flow through the intake valve undergoes alternate accelerations and decelerations, which require time. During the intake stroke, the mass of air moving through the passage to the intake valve is given velocity energy that may be converted to a slight pressure at the valve when the air mass still in the passage is stopped by its closure. Advantage of this phenomenon may be obtained at some engine speed to increase the air mass which enters the cylinder.

The engine speed at which the maximum volumetric efficiency is developed varies with the relative valve area, closure time, and other factors, including the diameter and length of the passage. The curves in **Fig. 6** show the characteristic falling off at high speeds from the inevitable throttling action as air flows at increased velocities through any restriction such as a valve or intake passage and venturi of a carburetor.

Volumetric efficiency has a direct effect on the mean effective pressure developed in a cylinder, on the torque, and on the power that may be realized at a given speed. Since power is a product of speed and torque, the peak power of an engine occurs at a higher speed than for maximum torque, where the rate of torque loss with any further increase in speed will exceed the rate of speed increase. An engine develops maximum power at a speed about twice that for maximum torque.

To obtain maximum torque and power, intake-valve closing may be delayed until the piston has traveled almost half the length of the compression stroke. At engine speeds below those where maximum torque is developed by this valve timing, some of the combustible charge that has been drawn into the cylinder on the intake stroke will be driven back through the intake valve before it closes. This reduces the effective compression ratio at wide-open throttle. The engine has an increasing tendency to develop spark knock as the speed and the resulting gas turbulence are reduced. *See* ENGINE MANIFOLD.

Supercharging spark-ignition engines. Volumetric efficiency and thus the mep of a four-stroke spark-ignition engine may be increased over a part of or the whole speed range by supplying air to the engine intake at higher than atmospheric pressure. This is usually accomplished by a centrifugal or rotary pump. The indicated power (power developed in the cylinder) of an engine increases directly with the absolute pressure in the intake manifold. Because fuel consumption increases at the same rate, the indicated specific fuel consumption (fuel flow rate per unit

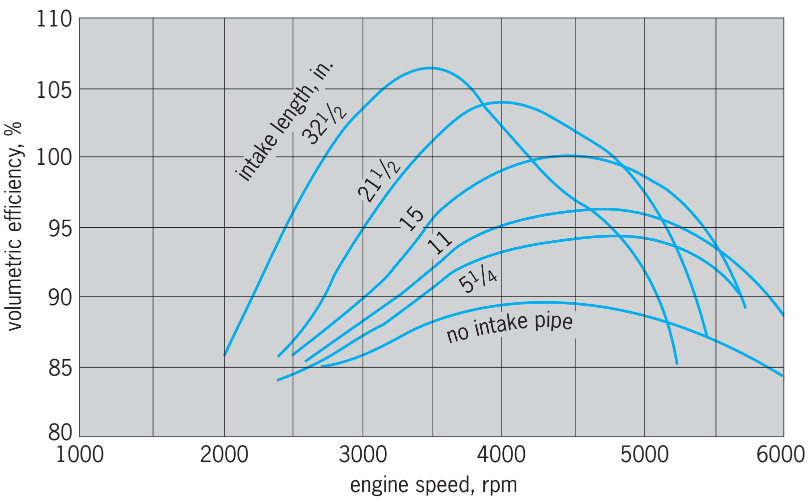


Fig. 6. Effect of intake-pipe length and engine speed on volumetric efficiency of one cylinder of a six-cylinder engine. 1 in. = 2.5 cm.

power output) is generally not altered appreciably by supercharging.

The three principal reasons for supercharging four-cycle spark-ignition engines are (1) to lessen the tapering off of mep at higher engine speed; (2) to prevent loss of power due to diminished atmospheric density, as when an airplane (with piston engines) climbs to high altitudes; and (3) to develop more torque at all speeds.

In a normal engine characteristic, torque rises as speed increases but falls off at higher speeds because of the throttling effects of such parts of the fuel intake system as valves and carburetors. If a supercharger is installed so as to maintain the volumetric efficiency at the higher speeds without increasing it in the middle-speed range, peak horsepower can be increased.

The rapid fall of atmospheric pressure at increased altitudes causes a corresponding decrease in the power of unsupercharged piston-type aircraft engines. For example, at 20,000 ft (6 km) the air density, and thus the absolute manifold pressure and indicated torque of an aircraft engine, would be only about half as great as at sea level. The useful power developed would be still less because of the friction and other mechanical power losses which are not affected appreciably by volumetric efficiency. By the use of superchargers, which are usually of the centrifugal type, sea-level air density may be maintained in the intake manifold up to considerable altitudes. Some aircraft engines drive these superchargers through gearing which may be changed in flight, from about 6.5 to 8.5 times engine speed. The speed change avoids oversupercharging at medium altitudes with corresponding power loss. Supercharged aircraft engines must be throttled at sea level to avoid damage from detonation or excessive overheating caused by the high mep which would otherwise be developed. *See* SUPERCHARGER.

Normally an engine is designed with the highest compression ratio allowable without knock from the fuel expected to be used. This is desirable for the highest attainable mep and fuel economy from an atmospheric air supply. Any increase in the volumetric efficiency of such an engine would cause it to knock unless a fuel of higher octane number were used or the compression ratio were lowered. When the compression ratio is lowered, the knock-limited mep may be raised appreciably by supercharging but at the expense of lowered thermal efficiency. There are engine uses where power is more important than fuel economy, and supercharging becomes a solution. The principle involved is illustrated in **Fig. 7** for a given engine. With no supercharge this engine, when using 93-octane fuel, developed an indicated mean effective pressure (imep; an average pressure forcing the piston down the cylinder) of 180 pounds per square inch (psi; 1240 kilopascals) at the borderline of knock at 8:1 compression ratio. If the compression ratio were lowered to 7:1, the mep could be raised by supercharging along the 7:1 curve to 275 imep before it would be knock-limited by the same fuel. With a 5:1 compression ratio it could be raised to 435 imep. Thus the imep could be raised until the cylinder became thermally limited by the temperatures of critical parts, particularly of the piston head.

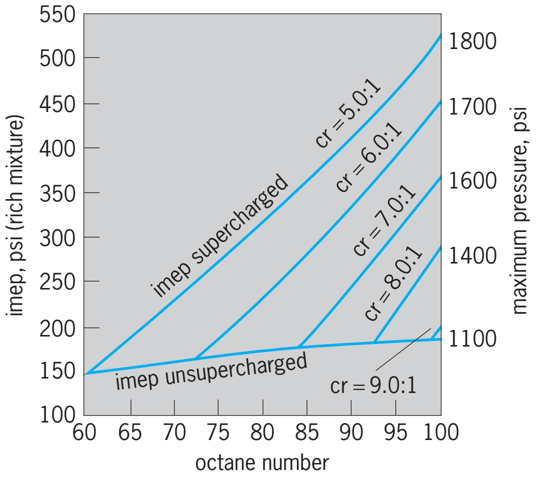


Fig. 7. Graph showing the relationship between compression ratio and knock-limited imep for given octane numbers, obtained by supercharging a laboratory engine. (*After H. R. Ricardo, The High-Speed Combustion Engine, 4th ed., Blackie, 1953*)

Engine balance. Rotating masses such as crank pins and the lower half of a connecting rod may be counterbalanced by weights attached to the crankshaft. The vibration which would result from the reciprocating forces of the pistons and their associated masses is usually minimized or eliminated by the arrangement of cylinders in a multicylinder engine so that the reciprocating forces in one cylinder are neutralized by those in another. Where these forces are in different planes, a corresponding pair of cylinders is required to counteract the resulting rocking couple.

If piston motion were truly harmonic, which would require a connecting rod of infinite length, the reciprocating inertia force at each end of the stroke would be as in Eq. (2), where W is the total weight

$$F = 0.000456WN^2s \tag{2}$$

of the reciprocating parts in one cylinder, N is the rpm, and s is the stroke in inches. Both F and W are in pounds. But the piston motion is not simple harmonic because the connecting rod is not infinite in length, and the piston travels more than half its stroke when the crankpin turns 90° from firing dead center. This distortion of the true harmonic motion is due to the so-called angularity a of the connecting rod, shown by Eq. (3), where r is the crank radius,

$$a = \frac{r}{l} = \frac{s}{2l} \tag{3}$$

s the stroke, and l the connecting rod length, all in inches.

Reciprocating inertia forces act in line with the cylinder axis and may be considered as combinations of a primary force—the true harmonic force from Eq. (2)—oscillating at the same frequency as the crankshaft rpm and a secondary force oscillating at twice this frequency having a value of Fa, which is added to the primary at top dead center and subtracted from it at bottom dead center. In general, harmonics above the second order may be disregarded. Therefore, for a connecting rod with the angularity $a = 0.291$, the inertia force caused by a piston at top dead center is about 1.29 times the pure harmonic force, and at bottom dead center it is about 0.71 times as large.

Where two pistons act on one crankpin, with the cylinders in 90° V arrangements, the resultant primary force is radial and of constant magnitude, and it rotates around the crankshaft with the crankpin. Therefore, it may be compensated for by an addition to the weight required to counterbalance the centrifugal force of the revolving crankpin and its associated masses. The resultant of the secondary force of the two pistons is 1.41 times as large as for one cylinder, and reciprocates in a horizontal plane through the crankshaft at twice crankshaft speed.

In four-cylinder inline engines with crankpins in the same plane, the primary reciprocating forces of the two inner pistons in cylinders 2 and 3 cancel those of the two outer pistons in cylinders 1 and 4, but the secondary forces from all pistons are added. Therefore, they are equivalent to the force resulting from a weight about $4a$ times the weight of one piston and its share of the connecting rod, oscillating parallel to the piston movement, having the same stroke, but moving at twice the frequency. A large a for this type of engine is advantageous. Where the four cylinders are arranged alternately on each side of a similar crankshaft, and in the same plane, both primary and secondary forces are in balance. Six cylinders in line also balance both primary and secondary forces.

V-8 engines use a crank arrangement in which the crankpins are in two planes 90° apart. Staggering the crankpins for pistons 1 and 2 90° from each other equalizes secondary forces, but the forces are in different planes. The couple this introduces is canceled by an opposite couple from the pistons operating on the crankpins for pistons 3 and 4.

Torsion dampers. In addition to vibrational forces from rotating and reciprocating masses, vibration may develop from torsional resonance of the crankshaft at various critical speeds. The longer the shaft for given bearing diameters, the lower the speeds at which these vibrations develop. On automotive engines, such vibrations are dampened by a viscous vibration damper or by a bonded-rubber vibration damper that is similar to a small flywheel coupled to the crankshaft through a rubber ring. The vibration damper may be combined with the pulley for an engine-accessory drive belt. *See* MECHANICAL VIBRATION.

Firing order. The firing order is the sequence in which the cylinders deliver their power impulses to the crankshaft. It is determined by such factors as engine design, ignition intervals, and crankshaft loading. Cylinder arrangements are generally selected for even firing intervals and torque impulses, as well as for balance.

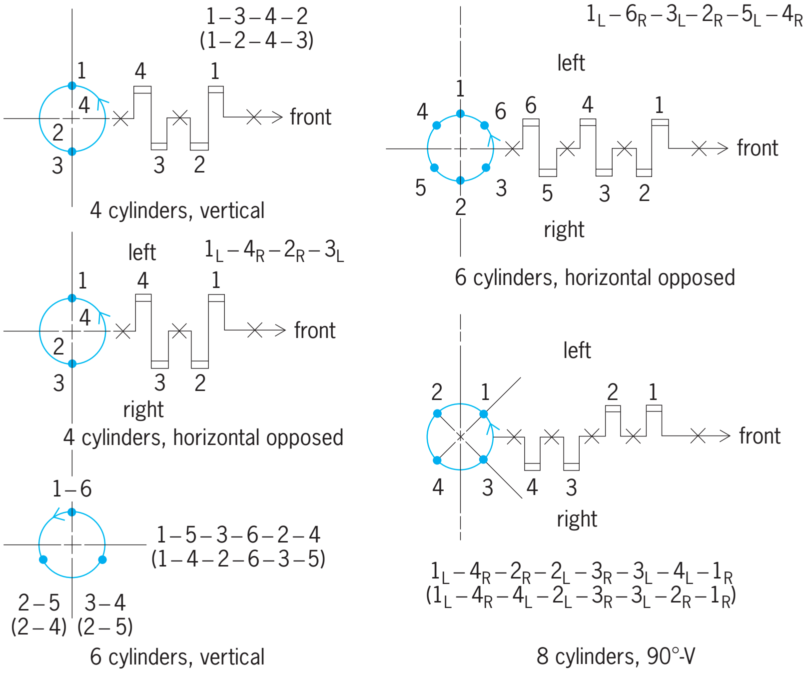


Fig. 8. Various cylinder arrangements and firing orders.

Figure 8 shows various cylinder arrangements and firing orders that have been used in automotive engines. The Society of Automotive Engineers (SAE) Standard for engine rotation and cylinder numbering provides that standard rotation is counterclockwise rotation of the crankshaft as viewed from the principal output end of the engine. If power can be delivered from either end, rotation shall be as viewed from the flywheel end.

Excluding radial engines or those with coplanar cylinder bore axes, cylinders may be numbered by either of two methods: (1) In single or multibank engines, the cylinders are numbered in the sequence in which the connecting rods are mounted along the crankshaft, beginning with the cylinder farthest from the principal output end. (2) In multibank engines, the cylinders may be numbered in sequence in each bank, starting with the cylinder farthest from the principal output end and designated right or left bank by suffixes R and L, such as 1R and 1L.

Cylinder bank and accessory locations are described as right or left when the engine is viewed from the flywheel or principal output end.

Compression-ignition engines.. In 1897, about 20 years after Otto first ran his engine, Rudolf Diesel successfully demonstrated an entirely different method of igniting fuel. Air is compressed to a pressure high enough for the adiabatic temperature to reach or exceed the ignition temperature of the fuel. Because this temperature is 1000°F (538°C) or higher, compression ratios of 12:1 to 23:1 are used commercially with compression pressures from about 440 to 800 psi (3 to 5.5 megapascals). The fuel is injected into the cylinders shortly before the end of the compression stroke, at a time and rate suitable to control the rate of combustion.

Compression ratio and combustion. The idealized diesel engine cycle assumes combustion at constant pressure. Like the Otto cycle, thermal efficiency increases with compression ratio, but also varies with the amount of heat added (at the constant pressure) up to the cutoff point where the pressure begins to drop from adiabatic expansion. *See* DIESEL CYCLE; DIESEL ENGINE.

Fuel injection. Early diesel engines used air injection of the fuel to develop extremely fine atomization and a good distribution of the spray. But the need for injection air at pressures of about 1500 psi (10 MPa) required expensive and bulky multistage air compressors and intercoolers.

A simpler fuel-injection method was introduced by James McKechnie in 1910. He atomized the fuel as it entered the cylinder by use of high fuel pressure and suitable spray nozzles. After considerable development, it became possible to atomize the fuel sufficiently to minimize the smoky exhaust that had been characteristic of the early airless or solid-injection engines. By 1930, solid injection had become the generally accepted method of injecting fuel in diesel engines.

During the 1980s, electronically controlled fuel injection began replacing the mechanical system. Eletronically controlled mechanically actuated unit injectors allowed injection pressure of 22,000 psi (150 MPa). *See* FUEL INJECTION.

Supercharged diesel engines. Combustion in a four-stroke diesel engine is improved by supercharging. Fuels that would smoke heavily and misfire at low loads will burn otherwise satisfactorily with supercharging. The indicated mean effective pressure rises directly with the supercharging pressure, until it is limited by the rate of heat flow from the metal parts surrounding the combustion chamber, and the resulting temperatures.

When superchargers of either the centrifugal or positive-displacement type are driven mechanically by the engine, the power required becomes an additional loss to the engine output. There is a degree of supercharge for any engine that develops maximum efficiency. A supercharge that is too high absorbs more power in the supercharger than is gained by the engine, especially at low loads. Another means of driving the supercharger is by an exhaust turbine, which recovers some of the energy that would otherwise be wasted in the exhaust. This may be accomplished with so small an increase of back pressure that little power is lost by the engine. The result is an appreciable increase in efficiency at loads high enough to develop the necessary exhaust pressure. *See* TURBOCHARGER.

Supercharging a two-cycle diesel engine requires some means of restricting or throttling the exhaust to build up cylinder pressure at the start of the compression stroke, and is used on a few large engines. Most medium and large two-stroke diesel engines are usually equipped with blowers to scavenge the cylinders after the working stroke and to supply the air required for the subsequent cycles. These blowers, in contrast to superchargers, do not build up appreciable pressure in the cylinder at the start of compression. If the capacity of such a blower is greater than the engine displacement, it will scavenge the cylinder of practically all exhaust products, even to the extent of blowing some air out through the exhaust ports. Such blowers, like superchargers, may be driven by the engine or by exhaust turbines.

Contrast between diesel and Otto engines. There are many characteristics of the diesel engine which are in direct contrast to those of the Otto engine. The higher the compression ratio of a diesel engine, the less the difficulties with ignition time lag. Too great an ignition lag results in a sudden and undesired pressure rise which causes an audible knock. In contrast to an Otto engine, knock in a diesel engine can be reduced by use of a fuel of higher cetane number, which is equivalent to a lower octane number. *See* CETANE NUMBER.

The larger the cylinder diameter of a diesel engine, the simpler the development of good combustion. In contrast, the smaller the cylinder diameter of the Otto engine, the less the limitation from detonation of the fuel.

High intake-air temperature and density materially aid combustion in a diesel engine, especially of fuels having low volatility and high viscosity. Some engines have not performed properly on heavy fuel until provided with a supercharger. The added compression of the supercharger raised the temperature and, what is more important, the density of the combustion air. For an Otto engine, an increase in either the air temperature or density increases the tendency of the engine to knock and therefore reduces the allowable compression ratio.

Diesel engines develop increasingly higher indicated thermal efficiency at reduced loads because of leaner fuel-air ratios and earlier cutoff. Such mixture ratios may be leaner than will ignite in an Otto engine. Furthermore, the reduction of load in an Otto engine requires throttling, which develops increasing pumping losses in the intake system. Neil MacCoull; Donald L. Anglin

Bibliography. *Bosch Automotive Handbook*, 1986; W. H. Crouse, *Automotive Engines*, 8th ed., 1994; J. B. Heywood, *Internal Combustion Engine Fundamentals*, 1988; L. C. R. Lilly (ed.), *Diesel Engine Reference Book*, 1984; K. Newton, W. Steeds, and T. K. Garrett, *The Motor Vehicle*, 1989; Society of Automotive Engineers, *SAE Handbook*, 4 vols., annually; C. F. Taylor, *The Internal Combustion Engine in Theory and Practice*, vols. 1 and 2, 1985.

Internal energy

A characteristic property of the state of a thermodynamic system, introduced in the first law of thermodynamics. For a static, closed system (no bulk motion, no transfer of matter across its boundaries), the change ΔU in internal energy for a process is equal to the heat Q absorbed by the system from its surroundings minus the work W done by the system on its surroundings. Only a change in internal energy can be measured, not its value for any single state. For a given process, the change in internal energy is fixed by the initial and final states and is independent of the path by which the change in state is accomplished.

The internal energy includes the intrinsic energies of the individual molecules of which the system is composed and contributions from the interactions among them. It does not include contributions from the potential energy or kinetic energy of the system as a whole; these changes must be accounted for explicitly in the treatment of flow systems. Because it is more convenient to use an independent variable (the pressure P for the system instead of its volume V), the working equations of practical thermodynamics are usually written in terms of such functions as the enthalpy $H = U + PV$, instead of the internal energy itself. *See* CHEMICAL THERMODYNAMICS; ENTHALPY.

Paul J. Bender

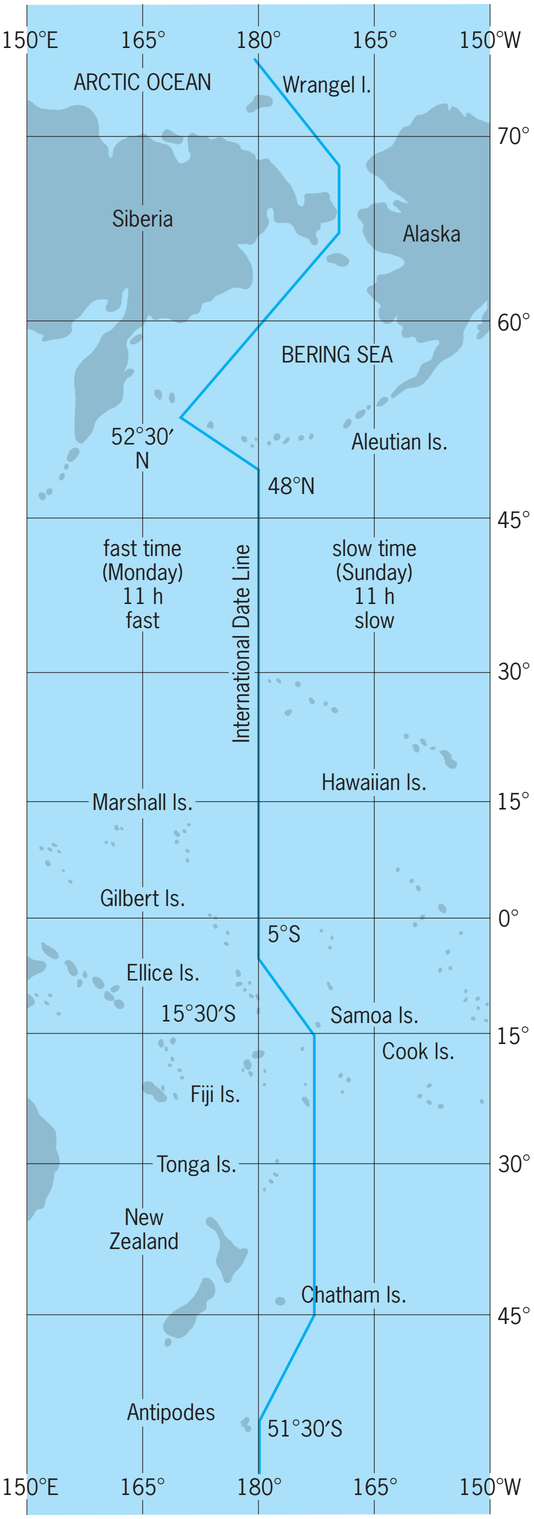
International Date Line

The 180° meridian, where each day officially begins and ends. As a person travels eastward, against the apparent movement of the Sun, 1 h is gained for every 15° of longitude; traveling westward, time is lost at the same rate. Two people starting from any meridian and traveling around the world in opposite directions at the same speed would have the same time when they meet, but would be 1 day apart in date. If there were no international agreement as to where each day should begin and end, there could be any number of places so designated. To eliminate such confusion, the International Meridian Conference, in 1884, designated the 180° meridian as the location for the beginning of each day. Thus, when a traveler goes west across the line, a day is lost; if it is Monday to the east, it will be Tuesday immediately as the traveler crosses the International Date Line. In traveling eastward a day is gained; if it is Monday to the west of the line, it will be Sunday immediately after the line is crossed.

An interesting example can be taken from conditions now nearly attainable with jet aircraft. If one could board such a plane, say at noon in Washington, D.C., and fly westward at that latitude around the world to return in 24 h, the rate would match the rotation of the Earth. Although constantly under the noontime Sun, this traveler would need to adjust her calendar 1 day ahead upon crossing the International Date Line, because she would arrive in Washington at noon, 24 h after embarking. Thus her calendar day would agree with that of the Washingtonians.

The 180° meridian is ideal for serving as the International Date Line (see **illus.**). It is exactly halfway around the world from the zero, or Greenwich, meridian, from which all longitude is reckoned. It also falls almost in the center of the largest ocean; consequently there is the least amount of inconvenience as regards population centers. A few deviations in the alignment have been made, such as swinging the line east around Siberia to keep that area all in the same day, and westward around the Aleutian Islands so that they will be within the same day as the rest of Alaska. Other minor variations for the same purpose have been made near the Fiji Islands, in the South Pacific. *See* MATHEMATICAL GEOGRAPHY.

Van H. English



The International Date Line.

Internet

A worldwide system of interconnected computer networks. The Internet has begun to transform the way that communication systems operate. The origins of the Internet can be traced to the creation of ARPANET (Advanced Research Projects Agency Network) as a network of computers under the auspices of the U.S. Department of Defense in 1969. Today, the Internet connects millions of computers around the world in a nonhierarchical manner unprecedented in the history of communications. The Internet is a product of the convergence of media, computers, and telecommunications. It is not merely a technological development but the product of social and political processes as well, involving both the academic world and the government (the Department of Defense). From its origins in a nonindustrial, noncorporate environment and in a purely scientific culture, it has quickly diffused into the world of commerce. While the Internet has had a sudden and dramatic impact on the global economic and social order, it took almost 30 years to emerge as a major technological force (*see* **table**).

The Internet is a combination of several media technologies and an electronic version of newspapers, magazines, books, catalogs, bulletin boards, and much more. This versatility gives the Internet its power. However, it is difficult to make precise predictions about the success of the Internet because of the complicated relationships it has created among technologies, markets, and political systems. Contributing to the development of the Internet were a wide variety of users who adopted the Internet to communicate among themselves in a semiprivate way. These users were not bound by any customs and practices, but initiated a culture of communication that was flexible, quick, unbounded, and nonhierarchical, where there was no superior authority in control of the communication system and no standards to govern performance.

The Internet has grown dramatically since about 1988, although its impact was not felt until about 1995. By mid-2000, estimates of the number of connected computers were around 80 million.

The average user gets an Internet connection with the help of an Internet service provider (ISP). An Internet service provider is a commercial organization that provides an Internet dial-up account for the user. In order to utilize the service, the user needs to have a dial-up program, an e-mail program, and a Web browser program. These generally come as a standard package with the purchase of a computer.

Technological features. The Internet's Technological success depends on its principal communication tools, the Transmission Control Protocol (TCP) and the Internet Protocol (IP). They are referred to frequently as TCP/IP. A protocol is an agreed-upon set of conventions that defines the rules of communication. TCP breaks down and reassembles packets, whereas IP is responsible for ensuring that the packets are sent to the right destination.

Data travels across the Internet through several levels of networks until it reaches its destination. E-mail messages arrive at the mail server (similar to the local post office) from a remote personal computer connected by a modem, or a node on a local-area network. From the server, the messages pass through a router, a special-purpose computer ensuring that each message is sent to its correct destination. A message may pass through several networks to reach its destination. Each network has its own router that determines how best to move the message closer to its destination, taking into account the traffic on the network. A message passes from one network to the next, until it arrives at the destination network, from where it can be sent to the recipient, who has a mailbox on that network. *See* LOCAL-AREA NETWORKS; WIDE-AREA NETWORKS.

TCP/IP. TCP/IP is a set of protocols developed to allow cooperating computers to share resources across the networks. The TCP/IP establishes the standards and rules by which messages are sent through the networks. The most important traditional TCP/IP services are file transfer, remote login, and mail transfer.

Early history of the Internet

Year	Event
1969	ARPANET (Advanced Research Projects Agency Network) online, connecting four university computer centers using NCP (network control protocol)
1972	With 30 host computers in network, public demonstration at a computer conference in Washington, DC; Internetworking Group (INWG) formed to establish standard protocols
1977	THEORYNET on line, providing e-mail to over 100 non-defense-contracting computer researchers at the University of Wisconsin
1979	UNIX User Network started, to send technical messages, and later adapted for other discussion groups
1981	Corporate funding of BITNET, providing e-mail to some academics
1982	INWG establishes TCP/IP (Transmission Control Protocol/Internet Protocol) as the standard on ARPANET
1983	MILNET online, devoted to operational military nodes formerly on ARPANET
1984	Over 1000 host computers connected to the Internet
1986	NSFNET (National Science Foundation Network) online; management of regional nets by private firms begins
1988	Private networks and for-profit service providers establish links to the Internet for commercial users
1989	Over 100,000 host computers connected to the Internet
1991	World Wide Web software distributed by the European Laboratory for Particle Physics (CERN)
1993	Contract to manage various functions of NSFNET awarded to three corporations; a graphical browser of the Web (MOSAIC) distributed by the National Science Foundation
1995	Widespread popularity of the Internet, particularly e-mail and Web, extending to small businesses, homes, and schools around the world
1997	Internet becomes truly global

The file transfer protocol (FTP) allows a user on any computer to get files from another computer, or to send files to another computer. Security is handled by requiring the user to specify a user name and password for the other computer.

The network terminal protocol (TELNET) allows a user to log in on any other computer on the network. The user starts a remote session by specifying a computer to connect to. From that time until the end of the session, anything the user types is sent to the other computer.

Mail transfer allows a user to send messages to users on other computers. Originally, people tended to use only one or two specific computers. They would maintain "mail files" on those machines. The computer mail system is simply a way for a user to add a message to another user's mail file.

Other services have also become important: resource sharing, diskless workstations, computer conferencing, transaction processing, security, multimedia access, and directory services.

TCP is responsible for breaking up the message into datagrams, reassembling the datagrams at the other end, resending anything that gets lost, and putting things back in the right order. IP is responsible for routing individual datagrams. The datagrams are individually identified by a unique sequence number to facilitate reassembly in the correct order. The whole process of transmission is done through the use of routers. Routing is the process by which two communication stations find and use the optimum path across any network of any complexity. Routers must support fragmentation, the ability to subdivide received information into smaller units where this is required to match the underlying network technology. Routers operate by recognizing that a particular network number relates to a specific area within the interconnected networks. They keep track of the numbers throughout the entire process.

Internet connection speeds. In the conventional system, a stand-alone personal computer accesses the Internet through a dial-up telephone connection via a modem. This provides a temporary connection, known as a SLIP (single-line Internet protocol) or PPP (point-to-point protocol) connection, that enables full access to the Internet as long as the telephone connection is maintained. *See* MODEM.

Cable modem. A more advanced system is via the cable modem, which guarantees Internet service at much higher speeds. For example, a 2-magabyte file, which might take 9 min to download in the conventional system with a 28.8-kilobit-per-second telephone modem, can be downloaded in 10 s using a cable modem. The cable modem is faster because it uses the digital fiber-optic network, which is superior to the available conventional telephone system. Since the cable modem is always connected, there is no need to log in or dial up. However, the cable modem tends to be rather slow at peak hours when a number of subscribers try to send or receive data at the same time. In addition, the number of lines available via cable modem is limited, and is not likely to increase appreciably in the near future, despite a rapidly growing demand for higher speeds.

DSL. The digital subscriber line (DSL) offers a faster Internet connection than a standard dial-up connection or the cable modem connection. DSL uses the existing telephone line and in most cases does not require an additional line. Once installed, the DSL router provides the customer site with a continuous connection to the Internet without any fear that it will cause a busy signal to occur over the telephone line. In its various forms, DSL offers users a choice of speeds ranging from 32 kilobits per second to more than 50 megabits per second. Most Internet service providers offer symmetric DSL data services at a series of speeds so that customers can choose the rate that meets their specific business needs. Over any given link, the maximum DSL speed is determined by the distance between the customer site and the central office. At the customer premises, a DSL router or modem connects the DSL line to a local-area network or an individual computer. DSL takes existing voice cables that connect customer premises to the telephone company's central office and turns them into a high-speed digital link.

ADSL. Asymmetric digital subscriber line (ADSL) is a broadband communication technology designed for use on regular telephone lines. It is called asymmetric because more bandwidth is reserved for receiving data than for sending data. It has the ability to move data over telephone lines at speeds up to 140 times faster than the fastest analog modems available.

Domain Name System. The addressing system on the Internet generates IP addresses, which are usually indicated by numbers such as 128.201.86.290. Since such numbers are difficult to remember, a user-friendly system has been created known as the Domain Name System (DNS). This system provides the mnemonic equivalent of a numeric IP address and further ensures that every site on the Internet has a unique address. For example, an Internet address might appear as crito.uci.edu. If this address is accessed through a Web browser, it is referred to as a URL (Uniform Resource Locator), and the full URL will appear as http://www.crito.uci.edu.

The Domain Name System divides the Internet into a series of component networks called domains that enable e-mail (and other files) to be sent across the entire Internet. Each site attached to the Internet belongs to one of the domains. Universities, for example, belong to the "edu" domain. Other domains are gov (government), com (commercial organizations), mil (military), net (network service providers), and org (nonprofit organizations).

An Internet address is made up of two major parts separated by an @ (at) sign. The first part of the address (to the left of the @ sign) is the username, which usually refers to the person who holds the Internet account and is often the person's login name or in some way identifies him or her. The second part of the address (to the right of the @ sign) contains the hostname (which can refer to a server on a network), followed by the Internet address, which together identify the specific computer where the person has an Internet e-mail account. For example, the address jlin@dee.uxyz.edu can be diagrammed as shown in the **illustration**.

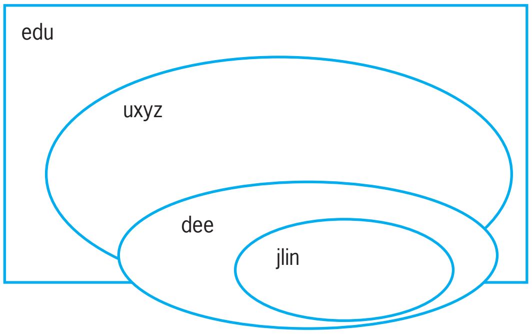


Diagram of the relationship of the sets of computers specified by the Internet address jlin@dee.uxyz.edu.

International domains. Because the Internet has become a global medium, many questions arise as to how the domains are to be managed and controlled, and by whom. Currently, a nongovernmental organization, the Internet Corporation for Assigned Names and Numbers (ICANN), performs a supervisory or advisory function in regard to standards, multilingual sites, domain names, and dispute resolution relative to international domain issues, and in particular, cybersquatting. Generally speaking, cybersquatting refers to the practice of registering a domain name of a well-known trademark in hopes that the owner of the trademark will eventually pay to get it back. Cybersquatting can also be considered an infringement of property right even if it does this unintentionally with the same effect. Since there is no case law or precedents governing Internet maintenance and management at the international level (or for that matter at the national level), many areas are open to question. Besides ICANN, other organizations specialize in various technical and management issues relating to Internet use and dissemination at the global as well as the local level. They work in consultation with ICANN.

Communication on the Internet. Electronic mail, or e-mail, may be the most heavily used feature of the Internet. Binary files, such as pictures, videos, sounds, and executable files, can be attached to e-mail messages. Because the Internet is not able to directly handle binary files in e-mail, the file must be encoded in one of a variety of schemes. Popular schemes are MIME and uuencode. The person who receives the attached binary file (called an attachment) must decode the file with the same scheme that was used to encode the file. Many e-mail software packages do this automatically. *See* ELECTRONIC MAIL.

Usenet. This is the world's largest electronic discussion forum, and it provides a way for messages to be sent among computers across the Internet. People participate in discussions on thousands of topics in specific areas of interest called newsgroups. There are at least 20 major hierarchies of newsgroups, such as recreation (identified by the letters "rec") and computers (identified by the letters "comp"). Within these major hierarchies are subcategories (such as rec.arts) and further subcategories (such as rec.arts.books).

Chats. One of the most immediate ways to communicate with others via the Internet is to participate in "chats" in which a person holds live keyboard "conversations" with other people on the Internet. A person can hold chats with many people simultaneo0usly.

Internet telephony. It is also possible to make telephone calls on the Internet. This can be done in two ways. A computer and special hardware and software can be used to make calls, so that communication is through the computer. In the second method, often referred to as Internet telephony, the user can make calls in the usual way, except that the call is routed over the Internet rather than the normal telephone service. This usually cuts the cost of making long-distance calls. In fact, the Internet telephone call is free; the user pays only for the Internet connection.

Intranets. An intranet is a packet of services that are accessible only within an organization. An intranet works over an organization's internal network and is set up for a wide variety of purposes, including e-mail, group brainstorming, group scheduling, access to corporate databases and documents, and videoconferencing, as well as buying and selling goods and services. Intranets use TCP/IP networks and technologies as well as Internet resources such as the World Wide Web, e-mail, Telnet, and FTP. An intranet is separated from the rest of the Internet by a firewall, a hardware and software combination that prohibits unauthorized access to the intranet. *See* COMPUTER SECURITY.

Telnet. A remarkable feature of the computer is that it lets an individual use the resources of a distant computer. A person can log onto another computer, issue commands as if he or she were at that computer's keyboard, and then gain access to all the computer's resources. This is done with an Internet resource called the Telnet. Telnet follows a client-server model, which allows a piece of software running on the user's own personal computer (the client) to use the resources of a distant server computer. The host allows many clients to access its resources at the same time; it is not devoted to a single user. *See* CLIENT-SERVER SYSTEM.

Many hosts on the Internet can be accessed by using Telnet. They are all different computers, so many of them do not work or look alike. For example, some might be UNIX-based systems, some might be Windows NT/2000®-based computers, some might be Macintoshes®, and so on. To connect using Telnet, the client must use terminal emulation, in order to ensure, in essence, that the client's keyboard and monitor function as the host expects. The most common terminal emulation is the VT-100, which is safe to use.

World Wide Web. The World Wide Web (WWW) is based on technology called hypertext. Most of the Web development has taken place at the European Laboratory for Particle Physics (CERN). The Web may be thought of as a very large subset of the Internet, consisting of hypertext and hypermedia documents. A hypertext document is a document that has a reference (or link) to another hypertext document, which may be on the same computer or in a different computer that may be located anywhere in the world. Hypermedia is a similar concept except that it provides links to graphic, sound, and video files in addition to text files.

In order for the Web to work, every client must be able to display every document from any server. This is accomplished by imposing a set of standards known as a protocol to govern the way that data are transmitted across the Web. Thus data travel from client to server and back through a protocol known as the HyperText Transfer Protocol (http). In order to access the documents that are transmitted through this protocol, a special program known as a browser is required, which browses the Web. *See* WORLD WIDE WEB.

Commerce on the Internet. Commerce on the Internet is known by a few other names, such as e-business, Etailing (electronic retailing), and e-commerce. The strengths of e-business depend on the strengths of the Internet. The Internet is available practically all over the world, at all times. It is simple to use, and the transaction costs for the end user are low. The costs are also extremely low for the vendors on the Internet, compared to traditional distribution channels. The Internet allows for two-way communication and is built on open standards. The two-way communication, in turn, allows for direct feedback of the customers, and the open standards mean interoperability between companies. It is fairly easy to integrate processes, services, and products, once they have been digitized.

Internet commerce is divided into two major segments, business-to-business (B2B) and business-to-consumer (B2C). In each are some companies that have started their businesses on the Internet, and others that have existed previously and are now transitioning into the Internet world. Some products and services, such as books, compact disks (CDs), computer software, and airline tickets, seem to be particularly suited for online business. Alladi Venkatesh

Bibliography. U. D. Black, *Internet Architecture: An Introduction to IP Protocols*, Prentice Hall PTR, Upper Saddle River, NJ, 2000; D. Comer, *The Internet Book: Everything You Need To Know about Computer Networking and How the Internet Works*, 2d ed., Prentice-Hall, Upper Saddle River, NJ, 3d ed., 2000; H. M. Deitel, P. J. Deitel, and T. R. Nieto, *Internet and the World Wide Web: How To Program*, Prentice Hall, Upper Saddle River, NJ, 2000; C. Dhawan, *Remote Access Networks: PSTN, ISDN, ADSL, Internet and Wireless*, McGraw-Hill, New York, 1998; E. A. Hall, *Internet Core Protocols: The Definitive Guide*, O'Reilly, Cambridge, MA, 2000; D. Minoli and A. Schmidt, *Internet Architectures*, Wiley, New York, 1999; M. Stefik, *The Internet Edge: Social, Legal, and Technological Challenges for a Networked World*, MIT Press, Cambridge, MA, 1999.

Interplanetary matter

Low-density dust or gas that fills the space in a planetary system around or between the planets. Most interplanetary matter in the inner solar system is dust created by collisions among asteroids or released by comets as they pass by the Sun. Ionized gas, launched at high speeds from the Sun as the solar wind, also permeates the solar system and creates a variety of important electromagnetic effects where it interacts with planets. Viewed from a nearby star, the interplanetary matter around the Sun would outshine the Earth. Recent detections of large quantities of similar material around other stars—many of which appear to harbor planets—have rekindled widespread interest in this subject.

Zodiacal dust. A cloud of dust, called zodiacal dust, fills the plane of the solar system interior to the asteroid belt. An observer on the Earth can see this dust with the unaided eye on some clear moonless nights because it scatters sunlight in the Earth's direction. It appears as a faint triangle of light in the ecliptic plane, which connects the constellations of the zodiac. Sometimes, backscattered sunlight from the dust can be seen as the Gegenschein, an additional faint patch of light in the antisolar direction. *See* ZODIACAL LIGHT.

Two satellites, the *Infrared Astronomical Satellite* (*IRAS*) and the *Cosmic Background Explorer* (*COBE*), made detailed maps of the zodiacal cloud as seen from Earth orbit, using detectors that were sensitive at a wide range of infrared wavelengths (1–200 micrometers). While a human being, who observes light at visible wavelengths, can hardly discern the zodiacal cloud among the stars and the galactic background, instruments in space that sense midinfrared wavelengths (7–30 μm) detect a bright glare from this dust cloud that dominates all other background sources. This glare is thermal emission akin to the glow from a red-hot stove. *See* HEAT RADIATION; INFRARED ASTRONOMY.

Most of this zodiacal dust comes from collisions among asteroids. Prominent families of asteroids, groups of asteroids that appear to have fragmented from a single larger body, create bands of dust that both *IRAS* and *COBE* detected. Comets also contribute large amounts of dust as they pass near the Sun. *IRAS* and *COBE* detected trails of dust left by individual comets, mixing into the background cloud. *See* ASTEROID; COMET.

The cloud dust takes the shape of a disk whose thickness is proportional to the distance from the Sun. At the asteroid belt, most of the dust sits closer than 0.2 astronomical unit from the midplane of the solar system. (One astronomical unit, or AU, equals the mean distance from the Earth to the Sun, 1.496×10^8 km or 9.3×10^7 mi.) Zodiacal dust has been measured within a few solar radii from the Sun in the F corona, where it reaches sublimation temperatures. In between the asteroid belt and the F corona, the number density of the particles falls off as $r^{-1.3}$, where r is the distance from the Sun. *See* SUN.

The sizes of the particles lie on a continuum from 1-μm-diameter dust particles to the size of asteroids, the largest of which is 950 km (590 mi) in diameter. The larger objects contain most of the interplanetary mass, but the smaller objects combined have most of the surface area, and so emit and reflect the most light. Near the Earth, the density of the cloud is about 3×10^{-23} g cm^{-3}.

A parameter called β, the ratio of the magnitude of the radiation pressure force on a particle to the magnitude of the Sun's gravitational force on the particle, determines the dynamical behavior of each particle. In general, small particles, which have a high mass-to-surface area ratio, have high β values. Particles smaller than about 1 μm tend to have β values near 1; radiation pressure rapidly blows them out of the solar system. These tiny particles that stream away from the Sun are called β-meteoroids. *See* RADIATION PRESSURE.

Solar photons add energy to the orbits of dust grains, but because the radiation pressure from these photons is outward, they do not add angular momentum. When the particles reradiate this energy, however, the radiation is beamed in the direction of the particle's motion, according to special relativity. The reaction to the force of the beamed radiation, the Poynting-Robertson drag, removes angular momentum from the orbit. Due to Poynting-Robertson drag, zodiacal dust with β somewhat less than 1 spirals in from the asteroid belt, reaching the Sun in roughly 10,000 years for a 10-μm particle, 10 times as long for a 100-μm particle. In general, because of the effects of radiation pressure and Poynting-Robertson drag, the steady presence of large amounts of dust in a planetary system implies the existence of a population of larger orbiting bodies to supply it. *See* POYNTING'S VECTOR.

The shape and composition of zodiacal dust particles vary depending on their origin, and none is homogeneous and spherical. However, these particles fly so fast through the inner solar system, up to 30 km s^{-1} (20 mi/s), that any attempt to catch them for close study destroys many particles completely, and biases the sample. One such biased sample, micrometeoroids collected at high altitudes in the Earth's atmosphere, consists of small particles that mostly resemble carbonaceous chondrites, the most primitive meteorites. They are aggregates of micrometer-sized silicate grains, with roughly solar composition, minus the volatile gases, plus some extra carbon and water. Particles that originate from comets are thought to be fluffy and porous, like a bird's nest; the spaces in the structure are where ice used to be before the comet neared the Sun and began to sublimate.

Interplanetary gas: the solar wind. The inner solar system also contains some interplanetary gas, or more specifically, plasma, most of which comes from the Sun in the form of the solar wind. The solar wind flows supersonically in all directions out of the corona of the Sun at typical speeds of 450 km s^{-1} (280 mi/s) and temperatures of 100,000 K. Like the corona, the solar wind consists mostly of ionized hydrogen. At the Earth's orbit, the density of the wind is roughly 5 particles/cm^3 (80 particles/in.^{3}), or about 8×10^{-24} g cm^{-3}. *See* SOLAR WIND.

Although the wind is electrically neutral, it supports a magnetic field with a strength of about 5 nanotesla (5×10^{-5} gauss) at the orbit of the Earth, which flows outward from the Sun along with the gas. Due to the rotation of the Sun, the path of the solar wind takes the shape of an archimedean spiral, and the magnetic field lines stretch so that they are locally parallel to this spiral. Interaction with the solar wind charges interplanetary dust particles to typical potentials of +5 volts. This charge allows the magnetic field carried by the wind to entrain the smallest β-meteoroids. This magnetic field also protects the solar system from low-energy galactic cosmic rays.

The zone near the Earth where the planet's magnetic field dominates the field in the solar wind is the Earth's magnetosphere. Mercury, Jupiter, Saturn, Uranus, and Neptune also have strong magnetic fields, and therefore large magnetospheres. The fields of these planets punch holes in the solar wind, and the solar wind sweeps these magnetospheres into long magnetotails stretching away from the Sun, removing ionized hydrogen and other gases from the atmospheres of these planets in the process. The interaction of the Earth's magnetic field and the solar wind causes the aurorae. *See* AURORA; JUPITER; MAGNETOSPHERE; MERCURY (PLANET); NEPTUNE; PLANETARY PHYSICS; SATURN; URANUS.

Comets also release gas, which sublimates from ice on their surfaces when they are near the Sun. This gas is mostly water and water-derived molecules such as OH and H_3O^+. However, in situ measurements of Comet Halley with the *Giotto* space probe suggest that comets may release up to six times as much mass in solids as in gas. The solar wind soon drags the gas released by comets into long plasma tails and finally pushes it out of the solar system.

The solar wind flows all the way out to about 50 AU, where the thermal pressure of the interstellar gas which surrounds the solar system overcomes the pressure of the solar wind, which is mostly ram pressure due to its high velocity flow. This interaction occurs at a shock front called the heliopause. Many other stars have winds and other outflows; these phenomena are not associated with the existence of planets. *See* INTERSTELLAR MATTER.

Outer solar system. The region between the asteroid belt and the orbit of Neptune appears to be relatively devoid of interplanetary dust. The *Pioneer 10* and *11* spacecraft and the *Ulysses* space probe found two minor populations of dust particles in this zone: the β-meteoroids ejected from the inner solar system, and a flow of interstellar dust grains on hyperbolic orbits. Beyond Neptune, a ring of comet-sized icy bodies, called the Kuiper Belt, orbits the Sun. Undoubtedly, Kuiper Belt Objects collide with one another and produce a disklike cloud of dust. Neptune's gravitational effects are expected to confine most of this dust to the region beyond 30 AU from the Sun. Dust so far from the Sun is cold, may be 50 K or less, so that it radiates most of its thermal energy at far-infrared wavelengths (30–200 μm). The low temperature makes the cloud too dim to detect from near the Earth, even though it may be just as massive as the zodiacal cloud. *See* KUIPER BELT; SPACE PROBE.

Other planetary systems. Other stars have circumstellar clouds of dust that may be analogous to the interplanetary dust in the solar system. The *Spitzer* satellite and other telescopes have imaged dusty disks around many nearby sunlike stars. Some of these stars are known to host planets as well as dust; the dust we see is interplanetary matter in these extrasolar planetary systems.

Surveys with the *Spitzer Space Telescope* have found many more examples of this phenomenon since its launch in 2003. The strong infrared emission is thought to come from starlight absorbed and reradiated by large quantities of circumstellar dust.

Circumstellar dust clouds around stars older than roughly 10 million years are often referred to as debris disks. Several stars in this age group with strong infrared emission, notably Beta Pictoris, AU Microscopii, HR 4796, Fomalhaut, Vega, Epsilon Eridani, and Tau Ceti, appear to be surrounded by rings or disks. These disks are thought to represent the "debris" left-over from the process of planet formation.

Some debris disks appear to harbor planetary systems. Some have planets that have been detected by close monitoring of the radial velocity of the star to look for the wobble caused by gravitational pull of an orbiting body. Other debris disks have planetary systems inferred from the structure asymmetries of the disks themselves. The debris disk around the star Beta Pictoris has been found to contain small amounts of gas, possibly arising from cometlike bodies. *See* EXTRASOLAR PLANETS.

The dust in the solar system emits and reflects at least six times as much light as the Earth. However, circumstellar dust around other stars is much more difficult to see than dust in the solar system. Known debris disks around other stars typically have hundreds or thousands of times as much dust as the solar system; less massive disks would be too faint to detect. Imaging disks often requires a special observational tool, like a coronagraph or an interferometer.

Debris disks often have relatively little dust interior to a distance of roughly 10 AU from their host stars. When a debris disk contains dust orbiting within a few astronomical units of the star, this dust is called exozodiacal dust, in analogy with the zodiacal dust in the solar system. Exozodiacal dust is considered one of the chief obstacles in the search for extrasolar planets that resemble the Earth.

Planets and interplanetary dust. The planets in the solar system interact with interplanetary dust in several important ways. For one, they can directly intercept some of the particles. Some 400 metric tons (4×10^5 kg, or 8×10^5 pounds-mass) of meteoritic material fall on the Earth every day. Objects more massive than about 1 kg (2.2 lbm) have enough material to survive the passage through the Earth's atmosphere and reach the ground; these are called meteorites. Particles less massive than 10^{-10} g (2×10^{-13} lbm), smaller than about 1 μm, can also survive the fall because they slow down so fast; these are called micrometeorites. Meteoroids larger than 1 g (2×10^{-13} lbm) produce visible trails as they disintegrate overhead; these are shooting stars (meteors). *See* METEOR; METEORITE; MICROMETEORITE.

Planets also affect interplanetary dust grains indirectly through their gravitational perturbations. For example, as zodiacal dust spirals past the Earth on its way into the Sun, some of the grains become trapped in resonant orbits with the Earth. If the number of years it takes for the particle to orbit the Sun is approximately a ratio of two small integers (like 3/2 or 5/3), the slight gravitational tugs the particle feels when it passes near the Earth tend to add up, reinforcing each other. The sum of all these tugs can temporarily counteract Poynting-Robertson drag and halt the dust grain's inward spiral. This process creates a ring of enhanced dust density around the Sun at a distance of roughly 1 AU. The *IRAS* and *COBE* missions detected both this ring and an asymmetry in the ring, a wake of further density enhancement trailing the Earth in its orbit.

No existing telescope can directly image a planet around another star. But the large-scale dynamical effects of a planet on a cloud of circumstellar dust are often easier to detect from afar than the planets themselves. For example, resonant phenomena like the Earth's dust ring may cause the asymmetries seen in maps of the dust around the nearby star Vega. Large planets may maintain some of the central holes in disks debris, just as Neptune probably prevents most dust generated in the Kuiper Belt from spiraling past it toward the Sun.

Marc J. Kuchner

Bibliography. E. Grün et al. (eds.), *Interplanetary Dust*, 2001; A. C. Levasseur-Regourd and H. Hasegawa (eds.), *Origin and Evolution of Interplanetary Dust*, 1991.

Interplanetary propulsion

Means of providing propulsive power for flight to the Moon or to a planet. A variety of different propulsion systems can be used. The space vehicles for these missions consist of a series of separate stages, each with its own set of propulsion systems. When the propellants of a given stage have been expended, the stage is jettisoned to prevent its mass from needlessly adding to the inertia of the vehicle. *See* ROCKET STAGING.

By expelling mass at high velocities, the propulsion systems provide impulse which permits the vehicle to execute its flight trajectory and the necessary maneuvers. Interplanetary or lunar trajectories, as well as the associated maneuvers, are complex, and different kinds of propulsion systems are desirable to meet the requirements of the various phases of flight. Although all propulsion systems actually used in interplanetary flight have been chemical rockets, several basically different propulsion systems have been studied. *See* ION PROPULSION; ROCKET PROPULSION; SPACE FLIGHT; SPACE NAVIGATION AND GUIDANCE; SPACE PROBE; SPACE TECHNOLOGY; SPACECRAFT PROPULSION; SPACECRAFT STRUCTURE.

George P. Sutton

Interpolation

A process in mathematics used to estimate an intermediate value of one (dependent) variable which is a function of a second (independent) variable when values of the dependent variable corresponding to several discrete values of the independent variable are known.

Suppose, as is often the case, that it is desired to describe graphically the results of an experiment in which some quantity Q is measured, for example, the electrical resistance of a wire, for each of a set of N values of a second variable υ representing, perhaps, the temperature of the wire. Let the numbers Q_i, $i = 1, 2, \ldots, N$, be the measurements made of Q and the numbers υ_i be those of the variable υ. These numbers representing the raw data from the experiment are usually given in the form of a table with each Q_i listed opposite the corresponding υ_i. The problem of interpolation is to use the above discrete data to predict the value of Q corresponding to any value of υ lying between the above υ_i. If the value of υ is permitted to lie outside these υ_i, the somewhat more risky process of extrapolation is used. *See* EXTRAPOLATION.

Graphical interpolation. The above experimental data may be expressed in graphical forms by plotting a point on a sheet of paper for each pair of values (υ_i, Q_i) of the variables. One establishes suitable scales by letting 1 in. represent a given number of units of υ and of Q. If υ is considered the independent variable, the horizontal displacement of the ith point usually represents υ_i and its vertical displacement represents Q_i.

If, for simplicity, it is assumed that the experimental errors in the data can be neglected, then the problem of interpolation becomes that of drawing a curve through the N data points P_i having coordinates (x_i, y_i) that are proportional to the numbers υ_i and Q_i, respectively, so as to give an accurate prediction of the value Q for all intermediate values of υ. Since it is at once clear that the N measurements made would be consistent with any curve passing through the points, some additional assumptions are necessary in order to justify drawing any particular curve through the points. Usually one assumes that the υ_i are close enough together that a smooth curve with as simple a variation as possible should be drawn through the points.

In practice the numbers υ_i and Q_i will contain some experimental error, and, therefore, one should not require that the curve pass exactly through the points. The greater the experimental uncertainty the farther one can expect the true curve to deviate from the individual points. In some cases one uses the points only to suggest the type of simple curve to be drawn and then adjusts this type of curve to pass as near the individual points as possible. This may be done by choosing a function that contains a few arbitrary parameters that may be so adjusted as to make the plot of the function miss the points by as small a margin as possible. For a more complete discussion of this topic *see* CURVE FITTING.

For many purposes, however, one uses a French curve and orients it so that one of its edges passes very near a group of the points. Having drawn in this portion of the curve, one moves the edge of the French curve so as to approximate the next group of points. An attempt is made to join these portions of the curve so that there is no discontinuity of slope or curvature at any point on the curve.

Tabular interpolation. This includes methods for finding from a table the values of the dependent variable for intermediate values of the independent variable. Its purpose is the same as graphical interpolation, but one seeks a formula for calculating the value of the dependent variable rather than relying on a measurement of the ordinate of a curve.

In this discussion it will be assumed that x_i and y_i ($i = 1, 2, \ldots, N$), which represent tabulated values of the independent and dependent variables, respectively, are accurate to the full number of figures given. Interpolation then involves finding an interpolating function $P(x)$ satisfying the requirement that, to the number of figures given, the plot of Eq. (1) pass

$$y = P(x) \tag{1}$$

through a selected number of points of the set having coordinates (x_i, y_i). The interpolating function $P(x)$ should be of such a form that it is made to pass readily through the selected points and is easily calculated for any intermediate value of x. Since many schemes are known for determining quickly the unique nth degree polynomial that satisfies Eq. (1) at any $n + 1$ of the tabulated values and since the value of such a polynomial may be computed using only n multiplications and n additions, polynomials are the most common form of interpolating function.

If the subscripts on x and y are reassigned so that the points through which Eq. (1) passes are now (x_0, y_0), (x_1, y_1), ..., (x_n, y_n), the polynomial needed in Eq. (1) may be written down by inspection, and one has Eq. (2), where Eq. (3) applies. Equation (3) is

$$y = \sum_{k=0}^{n} \frac{L_k(x)}{L_k(x_k)} y_k \tag{2}$$

$$L_k(x) = \frac{(x - x_0)(x - x_1) \cdots (x - x_n)}{(x - x_k)} \tag{3}$$

Lagrange's interpolation formula for unequally spaced ordinates. Since $L_k(x)$ vanishes for all x_s in the set $x_0, x_1, \ldots, x_n$ except x_k, substituting $x = x_s$ in the right-hand side of Eq. (1) gives rise to only one nonzero term. This term has the value y_s, as required.

For $n = 1$ Eqs. (2) and (3) give rise to Eq. (4),

$$y = \frac{x - x_1}{x_0 - x_1} y_0 + \frac{x - x_0}{x - x_1} y_1 \tag{4}$$

whose plot is a straight line connecting the points (x_0, y_0) and (x_1, y_1). Such an interpolation is referred to as linear interpolation and is used in all elementary discussions of interpolation; however, another equivalent form of this equation, given below in Eq. (12), is more often used in these cases.

Suppose the table were obtained from the equation $y = f(x)$, in which $f(x)$ is some mathematical function having continuous derivatives of all order up to and including the $(n + 1)$th. It is then possible to obtain an accurate expression for intermediate values of $f(x)$ by adding to the right-hand side of Eq. (2)

a so-called remainder term, expression (5), where

$$\frac{(x-x_0)(x-x_1)\cdots(x-x_n)}{(n+1)!}f^{(n+1)}(\xi) \qquad (5)$$

$f^{(n+1)}(\xi)$ is the $(n+1)$th derivative of $f(x)$ at some point $x = \xi$ lying between the smallest and largest of the values $x_0, x_1, \ldots, x_n$. Since the value of ξ is not known, the remainder term is used merely to set an upper limit on the truncation error introduced by using Lagrange's interpolation formula.

If the ordinates are equally spaced, that is, $x_s = x_0 + sh$ where h is the interval of tabulation, Lagrange's interpolation formula simplifies considerably and may be written as Eq. (6), where n is the degree of interpolating polynomial that now passes through the $n+1$ points $(x_{-p}, y_{-p}), \ldots, (x_{n-p}, y_{n-p})$. Here p is the largest integer less than or equal to $n/2$, and the $A_{-p}(u), \ldots, A_{n-p}(u)$ are polynomials in the variable shown as Eq. (7). The polynomials of the variable in Eq. (7) have been tabulated as functions of u.

$$y = \sum_{k=-p}^{n-p} A_k(u)y_k = A_{-p}(u)y_{-p} + \cdots + A_0(u)y_0 + \cdots + A_{n-p}(u)y_{n-p} \qquad (6)$$

$$u = \frac{x-x_0}{h} \qquad (7)$$

Inverse interpolation. If the value of y is known and the value of the corresponding independent variable x is desired, one has the problem of inverse interpolation. Since the polynomials $A_k(u)$ in Eq. (6) are known functions of u, and the values of y and y_k are also known, the only unknown in this equation is u. Thus the problem reduces to that of finding a real root u of an nth-degree polynomial in the range $0 < u < 1$. Having found u, one may find x from Eq. (7). For a discussion of numerical methods for solving for such a root and for more information on interpolation *see* NUMERICAL ANALYSIS.

One may also perform an inverse interpolation by treating x as a function of y. Since, however, the intervals between the y_i are not equal it is necessary to employ the general interpolation formula of Eq. (2) with the xs and ys interchanged.

Round-off errors. In the tabulated values, round-off errors y_i resulting from the need to express the entries y_i of the table as finite decimals will cause an additional error in the interpolated value of y that must be added to the truncation error discussed before. The effect of these errors on the application of Lagrange's interpolation formula is seen by Eq. (6) to be a total error ϵ_T in y given by Eq. (8). Letting e be the smallest positive number satisfying the condition $e > |\epsilon_k|$ for all k, one knows from Eq. (8) that relation (9) holds. Since the sum of the $A_k(u)$ is equal to 1, the factor

$$\varepsilon_T = \sum_k A_k(u)\varepsilon_k \qquad (8)$$

$$|\varepsilon_T| \le \sum_k |A_k(u) \,||\, \varepsilon_k \le e\sum_k |A_k(u)| \qquad (9)$$

$$\sum_k |A_k(u)|$$

in Eq. (9) is usually not much larger than 2 or 3 and thus the interpolated value of y has about the same round-off error as the individual entries.

Use of finite differences. For some purposes it is more convenient to use an interpolating formula based not so much on the entries y_i of a central difference table as upon their differences:

x_{-2}	y_{-2}				
		$\delta y_{-3/2}$			
x_{-1}	y_{-1}		$\delta^2 y_{-1}$		
		$\delta y_{-1/2}$		$\delta^3 y_{-1/2}$	
x_0	y_0		$\delta^2 y_0$		$\delta^4 y_0$
		$\delta y_{1/2}$		$\delta^3 y_{1/2}$	
x_1	y_1		$\delta^2 y_1$		$\delta^4 y_1$
		$\delta y_{3/2}$		$\delta^3 y_{3/2}$	
x_2	y_2		$\delta^2 y_2$		
		$\delta y_{5/2}$			
x_3	y_3				
.	.				
.	.				
.	.				

Each difference $\delta^k y_s$ is obtained by subtracting the quantity immediately above and to the left of it from the quantity immediately below and to the left; thus, Eq. (10) can be written, where k and $2s$ are required to be integers. For example, $\delta y_{1/2} = y_1 - y_0$ and $\delta^2 y_0 = \delta y_{1/2} - \delta y_{1/2}$.

$$\delta^k y_s = \delta^{k-1} y_{s+1/2} - \delta^{k-1} y_{s-1/2}. \qquad (10)$$

An interesting property of a difference table is that, if y, the dependent variable tabulated, is a polynomial of the nth degree in x, its kth difference column will represent a polynomial of degree $n-k$. In particular, its nth differences will all be equal and all higher differences will be zero. For example, consider a table of cubes and the difference table formed from it by the rule given above:

x	$y = x^3$				
0	0				
		1			
1	1		6		
		7		6	
2	8		12		0
		19		6	
3	27		18		0
		37		6	
4	64		24		0
		61		6	
5	125		30		
		91			
6	216				

Most functions $f(x)$, when tabulated at a small enough interval $\Delta x = h$, behave approximately as polynomials and therefore give rise to a difference table in which some order of difference is nearly constant. Consider, for example, the difference table of log x, in which the third differences fluctuate between 7 and 9 times 10^{-7}:

Experimental data, if taken at small enough interval of the independent variable, would be expected

x	$y = \log x$	δy	$\delta^2 y$	$\delta^3 y$
1.00	0.0000 000			
		43 214		
1.01	0.0043 214		–426	
		42 788		8
1.02	0.0086 002		–418	
		42 370		9
1.03	0.0128 372		–409	
		41 961		8
1.04	0.0170 333		–401	
		41 560		7
1.05	0.0211 893		–394	
		41 166		
1.06	0.0253 059			

to exhibit much the same behavior as a mathematical function except for the presence of experimental error. The presence of the latter will cause the differences to have a more or less random fluctuation. The size of the fluctuation may, in fact, be used to indicate the number of significant figures in the data.

The constancy of the third differences for log x indicates that for the accuracy and the interval used, a third-degree polynomial may be employed as an interpolating function. Since such a polynomial is determined by the choice of four coefficients, one would expect the interpolation formula to involve four numbers derivable from the difference table. Thus the forward-interpolation formula of Gauss, Eq. (11), can be written. If terminated after the fourth

$$\begin{aligned} y = y_0 + u\delta y_{1/2} &+ \frac{1}{2}u(u-1)\delta^2 y_0 \\ &+ \frac{1}{3!}u(u^2-1)\delta^3 y_{1/2} \\ &+ \frac{1}{4!}u(u^2-1)(u-2)\delta^4 y_0 \\ &+ \frac{1}{5!}u(u^2-1)(u^2-4)\delta^5 y_{1/2} + \cdots \quad (11) \end{aligned}$$

term, it represents a third-degree polynomial in $u = (x - x_0)/h$, and hence in x. It involves the four constants y_0, $\delta y_{1/2}$, $\delta^2 y_0$, and $\delta^3 y_{1/2}$. Since any one of the entries in the y column may be chosen as y_0, the differences required are picked from a central difference table, for example, in relationship to this entry. The interpolating polynomial obtained passes through the four points (x_{-1}, y_{-1}), (x_0, y_0), (x_1, x_2), (x_2, y_2). In general, the interpolating polynomial will pass through only those points whose y coordinate is needed to form the differences used in the formula.

If one terminates the series in Eq. (11) after the second term, one obtains Eq. (12). This is the linear

$$y = y_0 + u\delta y_{1/2} = y_0 + u(y_1 - y_0) \qquad (12)$$

interpolation formula most often used when making a simple interpolation in a table.

There are a great variety of interpolation formulas, such as Gregory-Newton's, Stirling's, and Bessel's, that differ mainly in the choice of differences used to specify the interpolating polynomial.

Difference equations. Repeated application of Eq. (10) may be used to express any difference in terms of the tabulated values, for example, Eqs. (13). Expressed in a more general form Eqs. (13) become Eqs. (14). If one sets the second difference equal to

$$\begin{aligned} \delta y_{1/2} &= y_1 - y_0 \\ \delta^2 y_0 &= y_1 - 2y_0 + y_{-1} \end{aligned} \qquad (13)$$

$$\begin{aligned} \delta f(x) &= f\left(x + \frac{h}{2}\right) - f\left(x - \frac{h}{2}\right) \\ \delta^2 f(x) &= f(x+h) - 2f(x) + f(x-h) \end{aligned} \qquad (14)$$

zero, one obtains a so-called difference equation for $f(x)$. In general, a difference equation is any equation relating the values of $f(x)$ at discrete values of x.

Difference equations play much the same role in analytical work as differential equations. Because they can be interpreted in terms of only those values of $f(x)$ tabulated at some interval $\Delta x = h$, they are admirably adapted to numerical computations. For this reason most numerical solutions of differential equations involve approximating the equation by a suitable difference equation.

For ordinary differential equations the transformation to a difference equation can be made by replacing each derivative in the equation by the appropriate difference expression according to formula (15),

$$\begin{aligned} &f^{(2k)}(x) \to \frac{1}{h^{2k}}\delta^{2k} f(x)\, f^{(2k+1)}(x) \to \\ &\frac{1}{2h^{2k+1}}\left[\delta^{2k+1} f\left(x + \frac{h}{2}\right) + \delta^{2k+1} f\left(x - \frac{h}{2}\right)\right] \qquad (15) \end{aligned}$$

where $f^{(n)}(x)$ designates the nth derivative of $f(x)$. The difference equation resulting can then, as mentioned before, be used to express the relationship between the values of $f(x)$ at the discrete points $x_s = x_0 + sh$, $s = 0, 1, 2, \ldots, n$.

Partial difference equations. Suppose one chooses to specify a function of two variables $f(x, y)$ by giving its value at some regular array of points in the xy plane having coordinates $(x_m, y_n$. Then, in place of a linear partial differential equation for $f(x, y)$ one has a linear partial difference equation, Eq. (16), where

$$\sum_{s,\,t} A_s f(x_{i+s}, y_{j+t}) = g(x_i, y_j) \qquad (16)$$

$g(x, y)$ is a known function and i and j any of the set of integers for which the difference equation has significance. A difference equation in which some of the $f(x_{i+s}, y_{j+t})$ occur to a power other than the first is termed a nonlinear difference equation.

If one employs a square lattice makeup of the points, Eqs. (17), then Laplace's differential equation is approximated by difference equation (18), where, for simplicity, Eq. (19) holds.

$$x_m = x_0 + mh \qquad y_n = y_0 + nh \qquad (17)$$

$$f_{i+1,j} + f_{i,j+1} + f_{i-1,j} + f_{i,j-1} - 4f_{ij} = 0 \qquad (18)$$

$$f_{mn} = f(x_m, y_n) \qquad (19)$$

See GRAPHIC METHODS; LATTICE (MATHEMATICS).

Kaiser S. Kunz

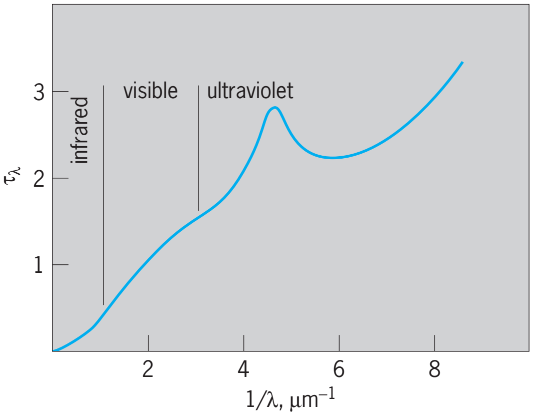
Bibliography. K. E. Atkinson, *Elementary Numerical Analysis*, 2d ed., 1993; R. L. Burden and

J. D. Faires, *Numerical Analysis*, 7th ed., 2001; J. Gregory and D. Redmond, *Introduction to Numerical Analysis*, 1994; J. Szabados and P. Vertesi (eds.), *Interpolation of Functions*, 1990.

Interstellar extinction

Dimming of light from the stars due to absorption and scattering by grains of dust in the interstellar medium. In the absorption process the radiation disappears and is converted into heat energy in the interstellar dust grains. In the scattering process the direction of the radiation is altered. Interstellar extinction produces a dimming of the light from stars situated beyond interstellar clouds of dust according to the equation $F_\lambda = F_\lambda(0)e^{-\tau_\lambda}$, where F_λ is the observed flux of radiation from the star at a wavelength λ, $F_\lambda(0)$ the flux that would be observed in the absence of interstellar extinction, and τ_λ the dimensionless optical depth for interstellar extinction at λ. Measures of the radiation from pairs of stars of similar intrinsic properties but with differing amounts of interstellar extinction can be used to obtain information about τ_λ, which can then be used to provide clues about the nature of the interstellar dust grains. *See* ABSORPTION OF ELECTROMAGNETIC RADIATION; SCATTERING OF ELECTROMAGNETIC RADIATION.

An interstellar extinction curve is a plot of the extinction optical depth τ_λ versus the wavelength or frequency of the radiation. It is common for astronomers to express extinction in the astronomical magnitude system. In that case the extinction in magnitudes, A_λ, and the extinction optical depth are related by $A_\lambda = 1.086\tau_\lambda$. A typical interstellar extinction curve for a 500-parsec path (1630 light-years, 9.58×10^{15} mi, or 1.542×10^{16} km) through the interstellar medium of the Milky Way disk is shown in the **illustration.** The spectral regions spanned by the curve are indicated. The extinction curve exhibits a nearly linear increase from infrared to visual wavelengths. In the ultraviolet there is a pronounced extinction peak near $1/\lambda = 4.6\ \mu m^{-1}$ or $\lambda = 217.5$ nm, followed by an extinction minimum and rapidly rising extinction to the shortest ultraviolet wavelengths for which data exist. Details not shown in the illustration include weak enhancements in extinction in the infrared near 20 and 9.7 μm and a very large number of such enhancements at visible wavelengths which are referred to as the diffuse interstellar features. A feature near 3.1 μm and other weaker infrared features appear in the extinction curves for sight lines passing through exceptionally dense interstellar clouds.



An interstellar extinction curve for a typical 500-parsec path through the interstellar medium of the Milky Way disk. Extinction optical depth τ_λ at wavelength λ is plotted versus $1/\lambda$ in μm^{-1}. The curve spans three segments of the spectrum.

The absorbing and scattering properties of solid particles depend on their size and composition. A detailed interpretation of the interstellar extinction curve of the illustration and other data relating to interstellar dust suggests that the interstellar grains of dust range in size from about 0.01 to 1 μm and are composed of silicate grains (to explain the 20- and 9.7-μm features) and probably some type of carbon grain (to explain the 217.5-nm extinction peak). The 3.1-μm feature in the extinction curve implies that the interstellar dust acquires coatings of water ice in the densest regions of interstellar space. Other ices, including those of ammonia (NH_3) and carbon monoxide (CO), are likely responsible for weaker infrared features observed in the densest interstellar clouds. A comparison of interstellar extinction with the absorption by interstellar atomic hydrogen reveals that the dust contains about 1% of the mass of the interstellar medium. *See* INTERSTELLAR MATTER.

Blair D. Savage

Bibliography. B. T. Draine, Interstellar dust grains, *Annu. Rev. Astron. Astrophys.*, 41:241–289, 2003; J. S. Mathis, Interstellar Dust and Extinction, *Annu. Rev. Astron. Astrophys.*, 28:37–70, 1990; D. C. B. Whittet, *Dust in the Galactic Environment*, 2d ed., 2002.

Interstellar matter

The gaseous material between the stars, constituting several percent of the mass of stars in the Milky Way Galaxy. Being the reservoir from which new stars are born in the Galaxy, interstellar matter is of fundamental importance in understanding both the processes leading to the formation of stars, including the solar system, and ultimately to the origin of life in the universe. Among the many ways in which interstellar matter is detected, perhaps the most familiar are attractive photographs of bright patches of emission-line or reflection nebulosity. However, these nebulae furnish an incomplete view of the large-scale distribution of material, because they depend on the proximity of one or more bright stars for their illumination. Radio observations of hydrogen, the dominant form of interstellar matter, reveal a widespread distribution throughout the thin disk of the Galaxy, with concentrations in the spiral arms. The disk is very thin (scale height 135 parsecs for the cold material; 1-pc is equal to 3.26 light-years, 1.92×10^{13} mi, or 3.09×10^{13} km) compared to its radial extent (the distance from the Sun to the galactic center is about 8000 pc, for example). Mixed in with the gas are small solid particles, called dust grains, of characteristic radius 0.1 micrometer. Although by mass the grains constitute less than 1% of the material, they have a pronounced effect through the extinction (absorption and scattering) of starlight. Striking examples of this obscuration are the dark rifts seen in the Milky Way, and the Coalsack Nebula in the southern sky. On average, the density of matter is only 15 hydrogen atoms per cubic inch (1 hydrogen atom per cubic centimeter; in total, 2×10^{-24} g · cm^{-3}), but because of the long path lengths over which the material is sampled, this tenuous medium is detectable. Radio and optical observations of other spiral galaxies show a similar distribution of interstellar matter in the galactic plane.

A hierarchy of interstellar clouds, concentrations of gas and dust, exists within the spiral arms. Many such clouds or cloud complexes are recorded photographically. However, the most dense, which contain interstellar molecules, are often totally obscured by the dust grains and so are detectable only through their infrared and radio emission. These molecular clouds, which account for about half of the interstellar mass, contain the birthplaces of stars. *See* GALAXY, EXTERNAL; MILKY WAY GALAXY.

Gas. Except in the vicinity of hot stars, the interstellar gas is cold, neutral, and virtually invisible. However, collisions between atoms lead to the production of the 21-cm radio emission line of atomic hydrogen. Because the Milky Way Galaxy is quite transparent at 21 cm, surveys with large radio telescopes have produced a hydrogen map of the entire Galaxy. Different emission regions in the Galaxy along the same line of sight orbit in the Galaxy with systematically different velocities relative to Earth and so are distinguishable by their different Doppler shifts. Supplemental information is obtained from 21-cm absorption-line measurements when hydrogen is located in front of a strong source of radio emission. These radio studies show that the gas is concentrated in clouds within the spiral arms, with densities typically 100 times the average of 15 atoms per cubic inch (1 atom per cubic centimeter). A typical size is 10 pc, encompassing the equivalent of 500 solar masses. These cold clouds (−316°F or 80 K) appear to be nearly in pressure equilibrium with a more tenuous warmer (14,000°F or 8000 K) phase of neutral hydrogen (whose mass totals that in clouds) and with even hotter ionized coronal-type gas (discussed below) in which they are embedded. *See* RADIO ASTRONOMY.

Other species in the gas are detected by the absorption lines they produce in the spectra of stars, and so are observable only in more local regions of the Milky Way Galaxy. Interstellar lines are distinguished from stellar atmospheric lines by their extreme narrowness and different Doppler shifts. High-dispersion spectra show that the lines are composed of several components possessing unique velocities which correspond to the individual clouds detected with the 21-cm line. Elements such as calcium, sodium, and iron are detected in optical spectra, but the more abundant species in the cold gas, such as hydrogen, carbon, nitrogen, and oxygen, produce lines only in the ultraviolet and require observations from satellites outside the Earth's atmosphere.

A broad ultraviolet line of O VI (oxygen atoms from which five electrons have been removed) has also been discovered; collisional ionization to such an extent requires a very hot gas ($\sim 3 \times 10^5$ K). Such hot gas emits soft x-rays; a Local Bubble of hot gas, extending to about 100 pc from the Sun, is a dominant contributor to the diffuse x-ray background radiation. A similar hot ionized component, called the coronal gas, occupies about half of the volume of interstellar space at the galactic midplane and extends significantly above the disk, but because of its low density (0.045 in.$^{-3}$ or 0.003 cm^{-3}) its contribution to the total mass density is small. *See* X-RAY ASTRONOMY.

When the Galaxy formed, only the elements hydrogen and helium (and traces of lithium) were present. In the course of the evolution of a star, heavier elements are built up by nuclear burning processes in the hot interior. Obviously, the interstellar gas will be gradually enriched in these heavy elements if some of this processed material can be expelled by the star. The most significant such event identified is the supernova explosion due to core collapse, which can occur late in the lifetime of a star whose original mass exceeds about 8 solar masses. A large fraction of the stellar mass is ejected, and in the course of the violent detonation further nucleosynthesis can occur. Also important are supernova explosions of white dwarfs and perhaps nova explosions on white dwarf surfaces. Computer simulations of the burning in supernova shocks predict remarkably large relative abundances of the isotopes. Lower-mass stars might be needed to explain the abundance of some elements such as carbon and nitrogen, the relevant mass loss being by stellar winds from red giants, culminating in the planetary nebula phase. *See* NOVA; NUCLEOSYNTHESIS; PLANETARY NEBULA; SUPERNOVA.

Through many cycles of star formation and mass loss from dying stars, the present interstellar element abundances have been built up to the levels seen in the atmospheres of relatively recently formed stars, including the Sun. The elements other than hydrogen and helium (principally carbon, nitrogen, and oxygen) constitute about 2% by mass. Direct examination of the cool interstellar gas using interstellar absorption lines reveals a considerable depletion of heavy elements relative to hydrogen. The atoms missing in the interstellar gas can be accounted for by the interstellar matter seen in solid particle form.

Molecules. In interstellar space, molecules form either on the surfaces of dust particles or by gas-phase reactions. In dense (greater than 1600 in.$^{-3}$ or 100 cm^{-3}) clouds where molecules are effectively shielded from the ultraviolet radiation that would dissociate them, the abundances become appreciable. These are called molecular clouds. Hydrogen is converted almost completely to its molecular form, H_2. The next most abundant molecule is carbon monoxide (CO), whose high abundance may be attributed to the high cosmic abundances of carbon and oxygen and the great stability of carbon monoxide.

The presence of molecules in interstellar space was first revealed by optical absorption lines. Unfortunately, most species produce no lines at optical wavelengths, and so the discovery of large numbers of molecules had to await advances in radio astronomy, for it is in this spectral band that molecular rotational transitions take place. Most lines are seen in emission. Dozens of molecules, with isotopic variants, are now known, the most widespread being carbon monoxide (CO), hydroxyl (OH), and formaldehyde (H_2CO). The most abundant molecule, H_2, has no transitions at radio wavelengths, but its presence in molecular clouds is inferred by the high density required for the excitation of the emission lines of other species and the lack of 21-cm emission from atomic hydrogen. In localized regions which have been heated by a shock or stellar wind to several thousand kelvins, the quadrupole infrared rotational-vibrational transitions of H_2 have been detected. Molecular hydrogen has also been observed in less opaque clouds by ultraviolet electronic transitions seen in absorption (carbon monoxide is also seen this way).

Molecular clouds. The existence of molecular clouds throughout the galactic disk has been traced by using the 2.6-mm emission line of carbon monoxide in a manner analogous to mapping with the 21-cm line. About half of the interstellar hydrogen is in molecular rather than atomic form, with enhancements in the inner 500 pc of the Galaxy and in a 3–7 kiloparsec ring. Molecular clouds exist primarily in the spiral arms, where they are thought to form behind the spiral shocks. The molecular clouds are quite cold ($-441°$F or 10 K), unless they have an internal energy source such as a newly forming protostar. They exist in complexes with the following typical characteristics: radius, 20 pc; density, 1.5×10^4 in.$^{-3}$ (1×10^3 cm^{-3}); mass, 10^5 solar masses. Within the complexes are more dense regions, whose counterparts are locally identified with the optically defined dark nebulae (discussed below). Only in the largest and most dense condensations can the rarer polyatomic (largely organic) molecules be seen. One of the chief such regions is the Orion Molecular Cloud OMC1, of which the Orion Nebula H II region is a fragment ionized by newly formed massive stars. The other region is the 5–10-pc core of the Sagittarius B2 cloud near the galactic center. Altogether, 114 molecules and molecular ions have been detected, many of which could be classed as prebiotic. It is interesting that the first stages of organic evolution can occur in interstellar space.

Unlike diffuse clouds, which are confined by the pressure of the surrounding medium, molecular clouds are confined by their own gravity. Molecular clouds are therefore susceptible to collapse into progressively denser phases, a sequence leading to the formation of new stars—usually in clusters of hundreds or thousands. *See* MOLECULAR CLOUD; ORION NEBULA; STAR FORMATION.

Masers. An intriguing phenomenon seen in some molecular emission lines is maser amplification to very high intensities. The relative populations of the energy levels of a particular molecule are determined by a combination of collisional and radiative excitation. The population distribution is often not in equilibrium with either the thermal gas or the radiation field; and if a higher-energy sublevel comes to have a higher population than a lower one, the electromagnetic transition between the states is amplified through the process of stimulated emission. Such nonequilibrium occurs in interstellar space in two distinct environments. *See* MASER.

The best-understood masers are the hydroxyl (OH), water (H_2O), and silicon monoxide (SiO) masers in the circumstellar envelopes of cool mass-losing red supergiant stars, because the relevant physical conditions relating to density, temperature, and radiation and velocity fields can be estimated independently from other observations. Positional measurements made possible by very long baseline interferometry show that the silicon monoxide and water masing regions are within a few stellar radii of the star, whereas the hydroxyl masers arise in less dense regions more than 100 stellar radii out. Even here the identification of the detailed pump (inversion-producing) mechanism is difficult, but it seems that the 1612-MHz hydroxyl maser is pumped by far-infrared emission by warm circumstellar dust grains and that the silicon monoxide masers are pumped by collisions.

The other masers are found in dense molecular clouds, in particular near compact sources of infrared and radio continuum emission identified as massive stars just being formed. Interstellar hydroxyl and water masers are widespread throughout the Galaxy, methyl alcohol (CH_3OH) masing is seen in OMC1, and silicon monoxide masers are seen in OMC1 and two other star-forming regions. The strongest hydroxyl masers are identified with underlying compact H II regions, probably occurring near the boundary. Although water masers are usually close to such compact H II regions, they are not coincident with the regions of radio or infrared continuum emission. Interstellar masers appear to require collisional pumping, which is thought to be due to powerful jets emitted from stars as they accrete. Peter G. Martin; Christopher D. Matzner

Chemical composition. At present, 123 interstellar molecular species are identified (see **table**). Mainly diatomic species are known, but recently a few more complex species have been seen in diffuse clouds (extinction $\lesssim 1$ magnitude); translucent clouds (1–5 magnitudes) contain molecules as complex as C_3H_2; over half the known species reside in cold dense clouds (>5 magnitudes); while nearly all species exist in warm dense star-forming cores in giant molecular clouds. The known molecules are largely organic. Only 21 are inorganic and stable. There are 36 stable organic species, and 67 unstable species, mostly organic, which include 33 free radicals, 14 (positive) ions, 4 simple rings, and 12 carbon chains as complex as $HC_{11}N$. Most of the unstable species were unknown terrestrially before their interstellar identification. The most rapidly growing list is that of the ions and radicals, because of new laboratory synthesis techniques.

Carbon chemistry is dominant, and long carbon-chain species are most prominent among organic

Known interstellar molecules

Inorganic species (stable)*

Diatomic		Triatomic	4-atom	5-atom
H_2	HCl	H_2O	NH_3	SiH_4^*
CO	HF	H_2S		
CS	$NaCl^*$	SO_2		
SiO	$AlCl^*$	OCS		
SiS	KCl^*	HNO		
PN	AlF^*	N_2O		
		SiH_2 ?		

Organic molecules (stable)*

Alcohols	Aldehydes and ketones	Acids	Hydrocarbons
CH_3OH (methanol)	H_2CO (formaldehyde)	HCN (hydrocyanic)	CH_2 (methylene)
CH_3CH_2OH (ethanol)	CH_3CHO (acetaldeyde)	$HCOOH$ (formic)	CH_4^* (methane)
CH_2CHOH (vinyl alcohol)	H_2CCO (ketene)	$HNCO$ (isocyanic)	$C_2H_2^*$ (acetylene)
$HOCH_2CH_2OH$ (ethylene glycol)	$(CH_3)_2CO$ (acetone)		$C_2H_4^*$ (ethylene)
	$HCCCHO$ (propynal)		C_6H_6 (benzene)?
	$HOCH_2CHO$ (glycol aldehyde)		

Amides	Esters and ethers	Organosulfur
NH_2CHO (formamide)	CH_3OCHO (methyl formate)	H_2CS (thioformaldehyde)
NH_2CN (cyanamide)	$(CH_3)_2O$ (dimethyl ether)	$HNCS$ (isothiocyanic acid)
NH_2CH_3 (methylamine)		CH_3SH (methyl mercaptan)

Paraffin derivatives	Acetylene derivatives	Rings	Other
CH_3CN (methyl cyanide)	CH_3N (cyanoacetylene)	C_2H_4O (ethylene oxide)	CH_2NH (methylenimine)
CH_3CH_2CN (ethyl cyanide)	CH_3C_2H (methylacetylene)		CH_2CHCN (vinyl cyanide)
	CH_3C_4H (methyldiacetylene)		$NaCN$ (sodium cyanide)

Unstable molecules*

Radicals		Ions	Rings	Carbon chains	Isomers
OH	C_2S	H_3^+	SiC_2	ℓ-C_3H	HNC
NO	$HCCN$	H_2D^+ ?	c-C_3H_2	C_3S	CH_3NC
NS	C_3^*	CH^+	c-C_3H	HC_5N	$HCCNC$
SO	C_3N	N_2H^+	SiC_3	HC_7N	$HNCCC$
SiN^*	C_5N	H_3O^+		HC_9N	$MgCN^*$
NH_2	C_3O	SO^+		$HC_{11}N$	HOC^+
CH	C_4H	CO^+		CH_3C_3N	
SH	C_5^*	HCO^+		CH_3C_5N ?	
CN	C_5H	HCS^+		C_4Si	
CSi^*	C_6H	$HOCO^+$		H_2CCC	
CP^*	C_7H	$HCNH^+$		H_2CCCC	
HCO	C_8H	CH_2D^+ ?		$H_2CCCCCC$	
CH_3	CH_2CN	HC_3NH^+			
C_2	$MgNC^*$	H_2COH^+			
C_2H	$ScCN$				
C_2O	$AlNC^*$				

* Molecules followed by asterisks are detected only in circumstellar envelopes (surrounding evolved, mass-losing, red giant stars), not yet in the interstellar medium.

compounds. Single, double, and triple carbon bonds are equally prevalent. No branched-chain species are known. Carbon-oxygen and carbon-nitrogen species are of comparable number, while nitrogen-oxygen bonds are rare. The molecules mostly contain carbon, oxygen, and nitrogen, reflecting the high cosmic abundances of these elements ($\sim 10^{-4}$ that of hydrogen). At $\sim 10^{-5}$ of hydrogen, sulfur and silicon are seen in fewer species, although organosulfur compounds (such as C_2S and C_3S) are noteworthy. Phosphorus, magnesium, aluminum, and sodium are seen in only one or two interstellar species, while those containing aluminum, chlorine, fluorine, and potassium arise also in unusual thermochemical equilibrium conditions in the innermost parts of circumstellar envelopes. Despite cosmic abundances similar to sulfur and silicon, the refractory elements magnesium and iron are underrepresented and are presumed locked in interstellar dust grains. *See* ELEMENTS, COSMIC ABUNDANCE OF; FREE RADICAL; ORGANOSULFUR COMPOUND.

Relative to H_2, into which virtually all hydrogen combines in molecular clouds, all other species occur only in trace amounts, ranging in abundance from carbon monoxide (CO) at 2×10^{-4} down to $HC_{11}N$ just detectable at 2×10^{-12}. The simplest species are usually the most abundant, although many complex species are surprisingly abundant also. As interstellar chemistry is far from equilibrium, abundances generally bear no relation to the constituent atomic abundances or to any thermodynamic properties.

Astrochemistry. Because of the very low temperatures and densities of the interstellar medium, molecules cannot form from atoms by normal terrestrial processes. Astrochemists have focused on three different chemistries. Most important is gas-phase chemistry, particularly involving molecular ions. Ion-molecule reactions satisfy the requirements of minimal activation energy and of rapid two-body rates. The (positive) ions are initiated by the cosmic-ray ionization of H_2, producing H_3^+, which then reacts with other abundant species such as carbon monoxide (CO) and N_2 to produce a large number of the observed species such as HCO^+ and N_2H^+. Ion fragments themselves react rapidly at low temperatures to form larger ions, which eventually recombine with electrons or neutralize by reaction with easily ionized metals. Slower reactions involving neutral atoms and molecules also produce several species. All of the observed molecular ions and a sizable number of the neutral species are successfully modeled by these processes, both in diffuse clouds where photochemistry is also important, and in cold and warm dense clouds opaque to ultraviolet radiation, which harbor many more species. Most of the observed patterns of species are explained. However, several complex molecules are not readily explained, nor is the mere existence of a gas phase in the cold dense clouds, given the rapid rate at which all molecules are expected to adsorb onto dust grains at low temperatures.

The chemistry of interstellar grains takes several forms. Grains can act as passive repositories of frozen molecular material while a cloud is in a cold dense phase, and the material later evaporates without chemical modification if the cloud warms as a result of nearby star formation. The same frozen material can undergo chemical reactions within the icy mantle, modifying the chemical composition before subsequent evaporation. Finally, molecules may be catalyzed from interstellar atoms arriving on dust grains. The ubiquitous H_2 molecule is the best example of a species that cannot form in the gas phase at sufficient rate but that catalyzes on grains and desorbs efficiently because of its high volatility. Details of the catalysis mechanism and the nature of the grain surfaces and composition are lacking, but experiments and models suggest many species can form readily on both ice-covered grains (cold clouds) and bare silicate or carbonaceous grains (warm clouds). In cold clouds especially, desorption is problematic, and the difficulty of desorption probably accounts for the lack both of refractory-element compounds and of several of the more complex species that also present problems for gas-phase formation. Recent searches for new species have targeted large complex stable molecules, such as are expected to form on grains. The two new alcohols and one new aldehyde suggest a chemistry that forms aldehydes and their corresponding reduced alcohols—for example, H_2CO/CH_3OH, long accepted as forming on grains from CO and H. Whether the next steps—CH_3CHO/CH_3CH_2OH or $HOCH_2CHO/HOCH_2CH_2OH$—can occur on grains is unknown. No known gas-phase formation is possible. It is believed that large molecules do form on grains, but desorb only by evaporation driven by nearby star formation, and then react in the gas phase with other gaseous species to form even larger molecules. *See* HETEROGENEOUS CATALYSIS.

Strong shocks abound in the interstellar medium, resulting from supernovae and expanding H II regions. These shocks briefly heat and compress the gas, producing required conditions for many high-temperature chemical reactions. The resulting shock chemistry has proved difficult to simulate. Models depend strongly on the type of shock assumed (dissociative or nondissociative, with or without a magnetic precursor) and do not yet provide a firm basis for comparison with observations. Certain highly abundant species such as hydroxyl (OH) and water (H_2O) in their maser forms, some inorganic sulfur species, and the refractory element species (SiO, SiS, PN) are the best candidates for shock formation.

Besides shocks and heating effects in star-forming regions, other processes also give rise to time dependence in interstellar chemistry. Ion-molecule chemistry in quiescent cold clouds predicts that the abundances of many species, especially the complex carbon-rich compounds (such as HC_3N and HC_5N), are up to 1000 times greater at ages of the order of 10^5 years than after the gas-phase chemistry reaches steady state ($\geq 10^7$ years), when essentially all carbon ends up as carbon monoxide (CO). Many clouds are therefore chemically young, implying a short lifetime, a dynamical mixing mechanism which renews the chemistry, or an overabundance of carbon. *See* SHOCK WAVE.

Isotopic composition. All of the most abundant isotopes of the elements occurring in interstellar molecules have been observed: H, D; ^{12}C, ^{13}C; ^{16}O, ^{18}O, ^{17}O; ^{14}N, ^{15}N; ^{32}S, ^{34}S, ^{33}S; ^{28}Si, ^{29}Si, ^{30}Si; and ^{25}Mg, ^{26}Mg. The relative abundances of the isotopically substituted molecules may differ from the nuclear isotopic ratios because of chemical fractionation, which favors the incorporation of heavier isotopes because of the reduced zero-point vibrational energy of the corresponding molecule. The effect is small for elements whose isotopic mass ratios are small, but becomes noticeable for $^{13}C/^{12}C$, and is enormous for deuterium/hydrogen (D/H), where the concentration factor favoring deuterium may approach 10^4, depending strongly upon the temperature and ionization fraction. Thus D/H cannot be determined from molecular observations; a value of $\sim 1.5 \times 10^{-5}$ is found for the local interstellar medium from ultraviolet observations of the atomic species. The other isotopic ratios can be found reliably from the molecules. Several of these ratios vary with distance from the galactic center.

Because the deuterium fractionation strongly increases with decreasing temperature, the deuterated ratio X_D/X_H has proved an important thermometer at the very lowest temperatures. Many multiply deuterated species have been recently detected, including

D_2CO, D_2H^+, D_2O, D_2S, ND_2H, D_3^+, ND^3. To fully replace all the H atoms with D atoms when D/H $\sim 1.5 \times 10^{-5}$, it appears that H_2D^+ is dramatically enhanced in gas depleted at least by 98% by molecules other than H_2.

Starting with the primordial elements hydrogen and helium, nuclear processing in stars forms the heavier elements and their isotopes, which are continuously cycled between stars and the interstellar medium. The concentrations of these elements and isotopes in the gas as a function of time depend on the rates of star formation, the stellar mass distribution, nuclear processes within the stars, and the rates of ejection of stellar material into the interstellar medium. The common isotopes ^{12}C and ^{16}O are formed first, and ^{13}C, ^{17}O, ^{18}O, ^{14}N, and ^{15}N are formed from them by secondary processes of stellar nucleosynthesis. All grow in abundance with time. For fast rates of stellar processing appropriate to the galactic center, the $^{12}C/^{13}C$ ratio is expected to decrease to a steady value of ~ 20 after 5×10^9 years. For slower rates of stellar processing characteristic of the galactic disk, $^{12}C/^{13}C$ should decrease more slowly, reaching ~ 90 when the Earth was formed 4.5×10^9 years ago, and ~ 70 at the present age of the Galaxy. An observed decrease in $^{18}O/^{17}O$ and a decrease in ^{15}N toward the galactic center are also consistent with advanced nucleosynthesis there. Stellar processing is not the only source of new elements: several isotopes of Li, Be, and B are produced primarily via spallation of interstellar C, N, and O nuclei by cosmic rays. *See* ISOTOPE. Barry E. Turner

Particles. The light of stars near the galactic disk is dimmed by dust grains, which both absorb the radiation and scatter it away from the line of sight. The amount of extinction at optical wavelengths varies approximately as the reciprocal of the wavelength, resulting in a reddening of the color of a star, much as molecular scattering in the Earth's atmosphere reddens the Sun, especially near sunrise and sunset. The dependence of extinction on wavelength is much less steep than it is for molecular scattering, indicating the solid particles have radii about 0.1 micrometer (comparable to optical wavelengths). Satellite observations show a continued rise in the extinction at ultraviolet wavelengths, which seems to require a size distribution extending to smaller interstellar grains. *See* INTERSTELLAR EXTINCTION.

By comparison of the observed color of a star with that predicted from its spectral features, the degree of reddening or selective extinction can be fairly accurately determined, but the total extinction at any given wavelength is more difficult to measure. On average, extinction by dust across 1000 pc in the galactic plane reduces a star's visual brightness by 80%. This requires the mass density of grains to be about 1% of the gas density. Since pure hydrogen or helium grains cannot exist in the interstellar environment, a major fraction of the heavier elements must be in the solid particles. The characteristic separation of micrometer-sized grains is about 8 m (13 km). Studies of reddening in conjunction with measurements of the 21-cm and ultraviolet Lyman-α lines of hydrogen and the 2.6-mm line of carbon dioxide show that dust and gas concentrations are well correlated. This correlation is borne out in infrared maps of diffuse 100-μm thermal emission from dust, made in the all-sky survey by the *Infrared Astronomical Satellite* (*IRAS*).

Because of the high concentration of interstellar material toward the galactic plane, it is extremely difficult to detect radiation with a wavelength less than 1 μm coming a large distance through the plane. Conversely, a line of sight to a distant object viewed out of the plane is much less obscured because the disk is so thin. The zone of avoidance, corresponding roughly to the area occupied by the Milky Way, is that region of the sky in which essentially no extragalactic object can be seen at optical wavelengths because of intervening dust. The dark rifts in the Milky Way result from the same obscuration. The component of starlight scattered rather than absorbed by the grains can be detected as a diffuse glow in the night sky near the Milky Way. However, it must be carefully separated from other contributions to the night sky brightness: the combined effect of faint stars, zodiacal light from dust scattering within the solar system, and airglow (permanent aurora). *See* AIRGLOW; AURORA; INTERPLANETARY MATTER; ZODIACAL LIGHT.

Interstellar dust particles are heated by the optical and ultraviolet interstellar radiation field and, being relatively cool, reradiate this absorbed energy in the infrared. Such thermal radiation accounts for a third of the bolometric luminosity of the Milky Way Galaxy. Measurements by *IRAS* at 100 and 60 μm are in accord with emission expected from the grains whose properties are deduced from extinction. On the other hand, the unexpectedly large portion (25%) at shorter wavelengths (25 and 12 μm) has been taken to indicate a significant population of very small ($\sim$0.001 μm) grains or large molecules. These have such small heat capacity that they can be heated transiently to a relatively high temperature by the absorption of a single photon.

The light of reddened stars is partially linearly polarized, typically by 1% but reaching 10% for the most obscured stars. The broad peak in polarization at yellow light, together with the correlation of the degree of polarization with reddening, suggests that the polarization and extinction are caused by the same dust grains. The grains must be both nonspherical and spinning about a preferred direction in space to produce polarization. The direction of this polarization is ordered across the sky, often parallel to the galactic plane. This is believed to be due to the organized galactic magnetic field. The field strength is very small; measurements based on the Zeeman effect in the 21-cm line of atomic hydrogen and the 18-cm line of hydroxyl indicate a few microgauss ($1\ \mu G = 10^{-10}$ tesla), with some compression in regions of higher density. Minute amounts (0.01%) of circular polarization have also been used to study the topology of the magnetic field.

The possible types of grain material can be restricted through considerations of relative cosmic abundances, but detailed identification is difficult because of the paucity of spectral features in the extinction curve. Silicates are suggested by an excess of absorption at 10 μm in front of strong infrared sources, and in the ultraviolet an absorption peak at 220 nanometers could be explained by a component of small graphite particles. Spatially extended red emission in excess of that expected from scattering in reflection nebulae has been interpreted as evidence for fluorescence of hydrogenated amorphous carbon solids. A popular theory of grain formation begins with the production of small silicate and carbon particles in the extended atmospheres of red supergiant stars. While in the circumstellar region, these grains are warmed by the starlight, so that they are detectable by their thermal emission on the near infrared (10 μm). Radiation pressure ejects these particles into the interstellar gas, where they become much colder (−433°F or 15 K). A dielectric mantle is then built up by accretion of the most abundant elements, hydrogen, carbon, nitrogen, and oxygen. Ices of water and carbon monoxide are detected by 3.1- and 4.67-μm absorption bands, respectively, but only deep within molecular clouds; these are probably in the form of volatile coatings accreted on more refractory grains.

Dark nebulae. A cloud of interstellar gas and dust can be photographed in silhouette if it appears against a rich star field; however foreground stars make distant clouds indiscernible. Many large dark nebulae, or groups of nebulae, can be seen in the Milky Way where the material is concentrated; these are coincident with molecular clouds. The distance to a dark nebula can be estimated by using the assumption that statistically all stars are of the same intrinsic brightness. When counts of stars within a small brightness range are made in the nebula and an adjacent clear region, the dimming effect of the cloud will appear as a sudden relative decrease in the density of stars fainter than a certain apparent brightness, which corresponds statistically to a certain distance. Alternatively, a lower limit to the distance is provided by the distance to the most distant unreddened stars in the same direction. One of the best-known and nearest dark nebulae is the Coal Sack, situated at a distance of 175 pc and visible in the Southern Hemisphere. Another example is the "Gulf of Mexico" area in the North America Nebula.

Obscuring clouds of all sizes can be seen against the bright H II regions described below. In many cases the H II regions and dark nebulae are part of the same cloud. The bay in the Orion Nebula is one such region, as well as the spectacular Horsehead Nebula. Even smaller condensations, called globules, are seen in the Rosette Nebula (**Fig. 1**) and NGC 6611 (M16; **Fig. 2**). The globules, which are almost completely opaque, have masses and sizes which suggest they might be the last fragments accompanying the birth of stars. *See* GLOBULE; NEBULA.

Bright nebulae. An interstellar cloud can also become visible as a bright nebula if illuminated by a nearby bright star. Whether an H II region or reflection nebula results depends on the quantity of ionizing radiation available from the star. To be distinguished from H II regions, but often also called bright gaseous nebulae, are shells of gas that have been ejected from stars. Included in this latter category are planetary nebulae and nova shells which have a bright emission-line spectrum similar to that of an H II region, and supernova remnants. *See* CRAB NEBULA.

H II regions. A star whose temperature exceeds about 45,000°F (25,000 K) emits sufficient ultraviolet radiation to completely ionize a large volume of the surrounding hydrogen. The ionized regions (Figs. 1 and 2), called H II regions, have a characteristic red hue resulting from fluorescence in which hydrogen, ionized by the ultraviolet radiation, recombines and emits the H ∝ line at 656.3 nm. Optical emission lines from many other elements have been detected, including the "nebulium" line of oxygen at 500.7 nm. H II regions are also conspicuous sources of free-free-radio emission characteristic of close

Fig. 1. Quadrant of the shell-shaped Rosette Nebula, showing dense globules of obscuring dust and gas silhouetted on the bright emission-line background of an H II region. The central hole is swept clear of gas by winds, supernovae, and ionization from newly formed massive stars; radiation pressure from the central star (lower left) acting on dust grains has also been suggested. The quadrant was photographed in red light with the 48-in. (122-cm) Schmidt telescope of the Palomar Observatory. (*California Institute of Technology/Palomar Observatory*)

Fig. 2. NGC 6611 (M16), a complex H II region in which the exciting stars are members of a cluster. Note the dark globules and elephant-trunk structures, and the bright rims where ionizing radiation is advancing into more dense neutral gas. The nebula was photographed in Hα + [N II] with the 200-in. (508-cm) telescope of the Palomar Observatory. (*California Institute of Technology/Palomar Observatory*)

electron-proton encounters in the 14,000°F (8000 K) gas. Some H II regions are seen only as radio sources because their optical emission is obscured by dust grains. Radio recombination lines of hydrogen, helium, and carbon, which result when the respective ions recombine to highly excited atoms, are also important.

An H II region can be extended with a relatively low surface brightness if the local density is low, as in the North America Nebula. However, the best-known regions, such as the Orion Nebula, are in clouds that are quite dense (1.5×10^4 to 1.5×10^5 H atoms per cubic inch, or 1×10^3 to 1×10^4 H atoms per cubic inch) compared to the average; dense clouds absorb ionizing radiation in a smaller region, and hence produce a brighter fluorescence. Since the brightest stars are also the youngest, it is not surprising to find them still embedded in the dense regions from which they formed.

Ionization of molecular gas increases its temperature from 30 K to 8000 K, and increases its pressure correspondingly. Enhanced pressures cause the H II region to explosively expand away from the forming star, often leading it to blowout through the side of the molecular cloud (the "blister phase," Fig. 2). The cloud mass unbound in this process is typically ten times that of the stellar association causing it, so H II regions are responsible for much of the inefficiency of star formation.

Reflection nebulae. In the absence of sufficient ionizing flux, the cloud may still be seen by the light reflected from the dust particles in the cloud. The scattering is more efficient at short wavelengths, so that if the illuminating star is white the nebula appears blue. The absorption or emission lines of the illuminating star appear in the nebular spectrum as well. Reflection nebulae are strongly polarized, by as much as 40%. Both the color and the polarization can be explained by dust grains similar to those which cause interstellar reddening. Extended near-infrared (2-μm) emission in some bright reflection nebulae provides evidence for thermal emission from transiently heated small grains. There are also numerous near-infrared spectral emission features consistent with C-H and C-C bending and stretching vibrations, and suggestive of polycyclic aromatic hydrocarbons (PAHs). Whether such compounds exist as free-flying molecules (or small grains), as coatings on larger dust grains, or both, is undecided.

Some reflection nebulae, such as those in the Pleiades, result from a chance encounter between a cloud and an unrelated bright star, providing a unique look at interstellar cloud structure. Other reflection nebulae appear to be intimately related to stars in early or late stages of stellar evolution. The Orion Nebula has an underlying reflection nebula arising from the dust in the gas cloud which produced the H II region. However, in this and other H II regions the emission-line radiation rather than the reflected light dominates the nebulosity.

Supernova remnants. Long after an exploding star has faded away, its supernova remnant (SNR) can still be seen. These are composed of gas expelled from the star itself and interstellar gas impelled into motion by it. Supernova remnants provide material to enrich the interstellar gas in heavy elements. The gas shell initially expands at up to 6000 mi · s^{-1} (10,000 km · s^{-1}); during this phase the supernova remnant is a strong emitter of radio waves through the synchrotron emission process in which internally accelerated relativistic electrons (cosmic rays) spiral in a magnetic field which has been amplified by turbulent and compressive motions. Later the supernova remnant is decelerated as it plows up more and more of the surrounding interstellar gas. Compression of the ambient magnetic field and cosmic-ray electron components of this gas leads again to a radio synchrotron source, with a characteristic shell structure. The supernova remnant generates a shock where it impinges on the interstellar medium, and because of the high relative velocity, temperatures of 10^5–10^6 K are reached in the post shock gas. This hot gas has been detected by its x-radiation in both emission lines and the free-free continuum. As the gas cools behind the shock, it produces optical emission lines too, giving rise to the bright shells seen in wide-field optical photographs (for example, the Veil Nebula which is part of the Cygnus Loop supernova remnant).

Supernovae caused by the collapse of a massive star's core produce a compact remnant, either a black hole or a neutron star (depending on the mass of the core before collapse). In the case of a magnetized, spinning neutron star, a pulsar will be observed within the supernova remnant. Pulsars emit a wind of relativistic particles and tightly wound magnetic fields, which illuminate the pulsar wind nebulae found within some supernova remnants via synchrotron emission (such as the Crab Nebula around the Crab Pulsar). Notably, no pulsar has been detected to date from the supernova 1987A in the Large Magellanic Cloud.

Supernova remnants provide an important source of energy for heating and driving turbulent motions in the cooler interstellar gas, and perhaps for accelerating cosmic rays. In low-density regions their influence can propagate through large regions of space, and they are probably responsible for maintaining the high-temperature coronal gas seen throughout the galactic disk. *See* COSMIC RAYS.

Magnetic fields and cosmic rays. The interstellar magnetic field, initially inferred from polarization of starlight, has been measured by means of the Zeeman splitting of the circularly polarized 21-cm emission, by the Faraday rotation of linearly polarized radio signals from pulsars and external galaxies, and from the synchrotron emission due to cosmic-ray particles spiraling about the field lines. The mean interstellar field strength is roughly 5 microgauss, or 10^{-5} of Earth's surface field. Magnetic field lines are efficiently coupled to charged particles, causing fields and gas to behave as a single fluid with properties of both components—notably, shearing of field lines engenders an opposing tension force. Magnetic forces contribute significantly to the support of gas against gravity, both above and below the midplane of the galaxy, and within individual molecular clouds where field strengths are 30 μG.

Cosmic rays are energetic charged particles—electrons, protons, and the nuclei of heavy elements—that pervade the interstellar medium. Cosmic rays observed at Earth include a low-energy variety produced by the Sun (which cause aurorae), interstellar particles of a wide range of energies (which cause genetic mutations and seed cloud formation) produced most likely by supernova remnants, and an extremely high energy population of mysterious extragalactic origin. Like interstellar gas, cosmic rays are effectively coupled to the magnetic field and affect the motion of the combined fluid. Indeed, cosmic-ray pressure is comparable to magnetic forces and turbulent motions in supporting gas away from the midplane. Cosmic rays contribute to the ionization and heating of the gas, but outside of dark clouds they are not as effective as the photoelectric absorption of far ultraviolet photons by dust grains. Their collisions with atomic nuclei produce light elements via spallation, and also produce π^0 mesons which decay to observable gamma-ray photons. *See* COSMIC RAYS.

Star formation. Superluminous stars such as those exciting H II regions cannot be very old because of the tremendous rate at which they are exhausting their supply of hydrogen for nuclear burning; the most luminous are under 100,000 years old. With this clear evidence for recent star formation, observations have been directed toward discovering stars even closer to their time of formation. Compact H II regions, such as the Orion Nebula, appear to be the first fragments of much larger molecular clouds to have formed stars. Ultracompact H II regions, seen in molecular clouds at radio wavelengths but totally obscured at optical wavelengths, appear to be an earlier stage in which the protostellar core has just become highly luminous. These are often called cocoon stars. Even earlier still are the compact infrared sources in which hot dust grains in a protostellar cloud are detected at wavelengths of 5–100 μm. These earliest phases are often associated with intense water and hydroxyl molecular maser emission. Examples of all of these stages are often found in the same region of space. In addition to the Orion molecular cloud, many regions such as the W3 radio and infrared sources associated with the visible H II region IC 1795 have been studied extensively.

Both the broad spectral energy distribution of thermal infrared emission and the spatial distribution of molecular emission from dense gas indicate that the final gravitational collapse of an interstellar cloud fragment commonly proceeds through an accretion disk. A flattened geometry is indeed expected from conservation of angular momentum during the collapse. In the protostar stage, radiant energy is derived from gravitational potential energy released as material accretes onto the star through the disk. A commonly associated phenomenon is the ejection of substantial amounts of gas in high-velocity bipolar outflows along in the direction perpendicular to the disk. These outflows are observed as optical emission lines from collimated jets and from shocked knots called Herbig-Haro objects, and as radio emission from carbon monoxide in extended lobes. Like H II regions, violent protostellar outflows are responsible for casting molecular gas away from forming stars and limiting the efficiency with which molecular clouds are converted into stars. *See* HERBIG-HARO OBJECTS; PROTOSTAR.

Although the details are far from understood, there is evidence that star formation is a bimodal process, with more massive stars forming only in some giant molecular cloud complexes which aggregate in the spiral arms of the Milky Way Galaxy, and less massive stars occurring independently over the whole range of molecular cloud sizes. Low-mass stars are less luminous and cooler and so do not produce the diagnostic compact H II regions. In the nearby Taurus molecular cloud, numerous dense (greater than 1.5×10^5 in.$^{-3}$ or 10^4 cm^{-3}) stellar mass cores, detected by ammonia (NH_3) emission, appear to be the immediate precursors of low-mass stars. This association is corroborated by their coincidence with other signposts of star formation: *IRAS* infrared sources, bipolar outflows, and T Tauri stars, which are low-mass pre-main-sequence stars.

Molecular clouds are supported against wholesale gravitational collapse by turbulent motions, probably associated with waves in the interstellar magnetic field, and are eventually dispersed by mechanical energy which results from the star formation process itself (bipolar outflows, expanding H II regions, stellar winds). On average, the star formation rate corresponds to a 2% conversion efficiency of molecular cloud material, but occasionally a higher efficiency, in excess of 30%, can lead to a gravitationally bound cluster of stars such as the Pleiades or NGC 6811 (M16; Fig. 2). Overall, the Milky Way Galaxy processes about five solar masses of interstellar matter per year into new stars, somewhat larger than the rate at which dying stars replenish the gas, so that the present interstellar medium is decaying on a time scale of about 1×10^9 years. A supply of fresh gas from outside the galactic disk is suggested by the fact that this is one-tenth of the Galaxy's age. *See* INFRARED ASTRONOMY; STELLAR EVOLUTION.

Peter. G. Martin; Christopher D. Matzner

Bibliography. W. B. Burton, B. G. Elmegreen, and R. Genzel, *The Galactic Interstellar Medium*, 1992; L. D'Hendecourt, C. Joblin, and J. Anthony (eds.), *Solid Interstellar Matter: The ISO Revolution*, Springer-Verlag, 1999; E. Levy, J. I. Lunine, and M. S. Matthews (eds.), *Protostars and Planets III*, 1993; V. Mannings, A. Boss, and S. Russell (eds.), *Protostars and Planets IV*, Unversity of Arizona Press, 2000; D. E. Osterbrock and G. J. Ferland, *Astrophysics of Gaseous Nebulae and Active Galactic Nuclei*, 2d ed., 2005; J. Palous, W. B. Burton, and P. O. Lindblad (eds.), *Evolution of Interstellar Matter and Dynamics of Galaxies*, 1992; B. E. Turner, Recent progress in astrochemistry, *Space Sci. Rev.*, 51:235–337, 1989; E. F. van Dishoeck and G. A. Blake, Chemical evolution of star-forming regions, *Annu. Rev. Astron. Astrophys.*, 36:317–368, 1998; T. L. Wilson and R. T. Rood, Abundances in the interstellar medium, *Annu. Rev. Astron. Astrophys.*, 32:191–226, 1994.

Intestine

The tubular portion of the digestive tract, usually between the stomach and the cloaca or anus. The detailed functions vary with the region, but are primarily digestion and absorption of food.

The structure of the intestine varies greatly in different vertebrates (see **illus.**), but there are several common modifications, mainly associated with increasing the internal surface area. One, seen in many fishes, is the development of a spiral valve; this turns the intestine into a structure resembling a spiral staircase. Another, seen in some fish and most tetrapods, is simply elongating and then coiling the intestine. This can reach extremes in large herbivores: Oxen have intestinal lengths of over 150 ft (45 m). In numerous forms there are blind pouches, or ceca, off part of the intestine. In fish these are commonly at the anterior end; in tetrapods they generally lie at the junction between the large and small intestines. In all vertebrates the inner surface of the intestine is irregular, with ridges and projections of various sorts; these reach their maximum development in the extremely fine and numerous finger-shaped villi found in mammals.

In humans the intestine consists of the small and large intestines. The small intestine is further divided into three major parts: the duodenum, the jejunum, and the ileum. The duodenum, 10–12 in. (25–30 cm) long, begins at the pyloric sphincter of the stomach and curves around the head of the pancreas on the right side of the anterior part of the abdomen. It receives the ducts of the biliary system and the pancreas. The jejunum and ileum are about 19 ft (6 m) long and form a much-coiled tube that empties at right angles into the large intestine through the ileocolic valve. The large intestine, or colon, consists of five parts: the ascending, transverse, descending, and sigmoid regions, and the terminal rectum which empties into the anal canal.

The microscopic structure of the intestine comprises an inner glandular mucosa, a muscular coat, and an outer serosa of connective tissues which is covered in most areas by peritoneum.

The intestine is supported by dorsal mesenteries of varying extent, which contain an extensive system of arteries, veins, lymphatics, and nerves to the various regions. *See* DIGESTIVE SYSTEM.

Thomas S. Parsons

Intra-Americas Sea

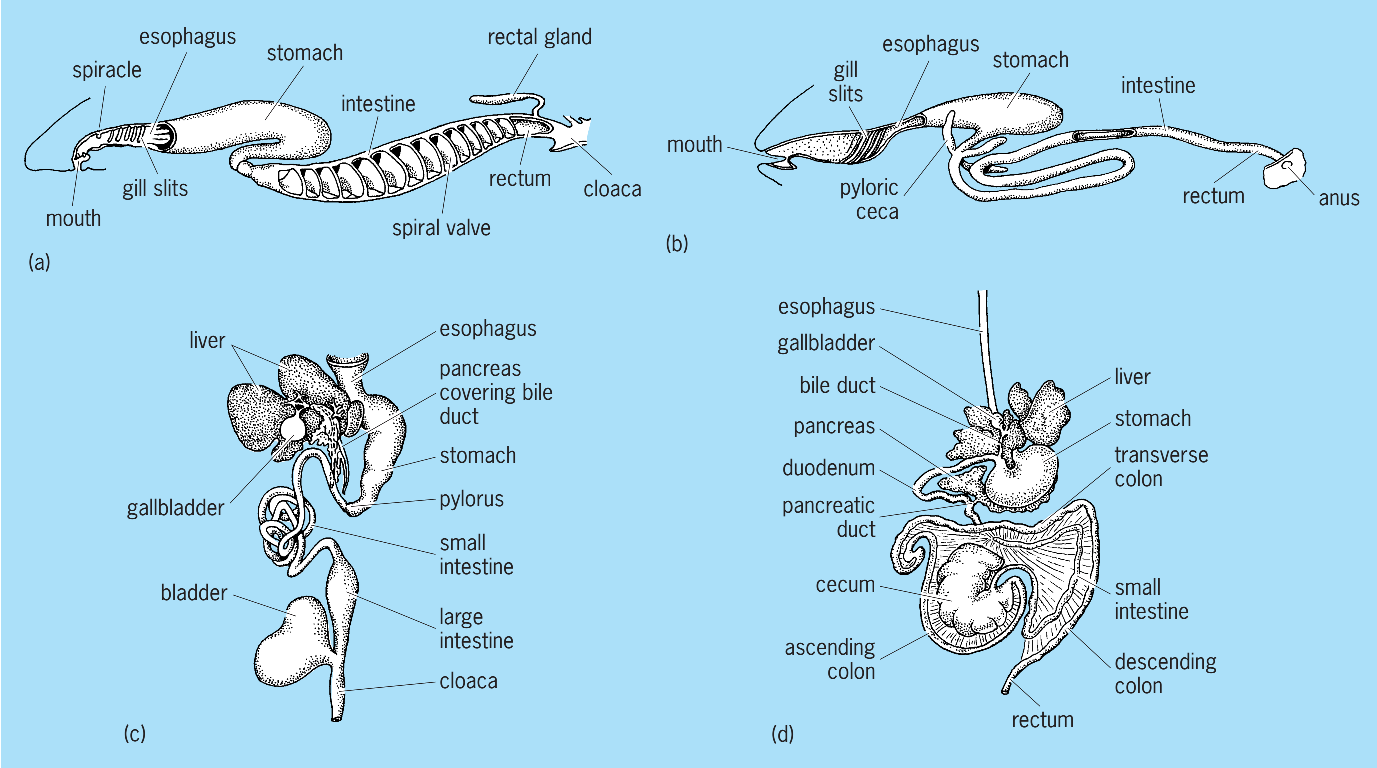
That area of the tropical and subtropical western North Atlantic Ocean encompassing the Gulf of Mexico, the Caribbean Sea, the Bahamas and Florida, the northeast coast of South America, and the juxtaposed coastal regions, including the Antillean Islands.

Meteorology. Meteorologically, the Intra-Americas Sea (IAS) is a transition zone between truly tropical conditions in the south and a subtropical climate in the north. The annual variability of the intertropical convergence zone (ITCZ) controls the easterly direction (that is, from the east) of the trade winds, where low surface pressures and rising air create the characteristic cumulus and cumulonimbus clouds. Except for winter when frontal passages bring westerly winds to the northern Intra-Americas Sea, the intertropical convergence zone dominates the weather. The Intra-Americas Sea is also the region that either spawns or interacts with the intense tropical storms known locally as the West Indian Hurricane. Typically there are 9 tropical storms per year, six of which become intense enough (74-mi/h or 119-km/h surface winds) to be classified as hurricanes. These storms are destructive because of their winds and storm surge, but they provide significant amounts of the annual rainfall to many of the Intra-Americas Sea islands and coastal zones. Air flowing over the Intra-Americas Sea acquires moisture that is the source of much of the precipitation over the central plains of North America. Its quality has caused some meteorologists to give it the name American Monsoon. *See* HURRICANE; MONSOON METEOROLOGY; TROPICAL METEOROLOGY.

Circulation. Ocean currents of the Intra-Americas Sea are dominated by the Gulf Stream System. Surface waters flow into the Intra-Americas Sea through the passages of the Lesser Antilles, and to a lesser extent through the Windward Passage between Cuba and Haiti, and the Anegada Passage between Puerto Rico and Anguilla. These inflowing waters form the Caribbean Current, which flows westward and northward into the Gulf of Mexico through the Yucatán Channel, where it is called the Yucatán Current. Here it is clearly recognized as a western boundary current because of its high surface speed, depth, volume transport, and very clear waters.

As the water passes through the Gulf of Mexico, it is known as the Gulf Loop Current because it flows far to the north, loops anticyclonically (clockwise) and flows south along the continental slope off Florida, and finally loops again cyclonically as it turns to exit into the Straits of Florida. In the Straits of Florida the current is called the Florida Current, a name it retains until about the latitude of Cape Canaveral, after which is it usually called the Gulf Stream. The Gulf Loop Current forms large (186-mi or 300-km diameter) anticyclonic eddies at an average rate of every 10–11 months; these eddies drift into the western Gulf of Mexico, spinning down and releasing their momentum, heat, and salt into the basin. Unlike the European Mediterranean whose deep waters are formed locally, the subsurface waters of the Intra-Americas Sea can be traced upstream as far south as Antarctica (Antarctic Intermediate Water) and as far east as the Canary Islands (Subtropical Underwater). Deep water flowing southward along the eastern boundary of the Intra-Americas Sea is known as the Deep Western Boundary Current, which has its source in the Arctic. Thus the Intra-Americas Sea's oceanic circulation is truly part of a global pattern. *See* GULF OF MEXICO; GULF STREAM; OCEAN CIRCULATION.

Surrounding land areas. Land surrounding the Intra-Americas Sea is often fringed by coral reefs, and there are five tectonic plates that influence its geophysics: the North American, South American,



Vertebrate digestive tracts, showing structure of fish, amphibian, and mammalian intestines. (*a*) Shark. (*b*) Perch. (*c*) Frog. (*d*) Guinea pig. (*After A. S. Romer*, *The Vertebrate Body*, *3d ed.*, *Saunders*, *1962*)

Caribbean, Cocos, and Nazca plates. Earthquakes are not uncommon, nor are the associated seismic sea waves (tsunamis) that often cause extensive coastal destruction but that are usually considered to be only Pacific Ocean phenomena. The coastal lands are often the site of extensive mangrove forests; the coastal waters are highly productive in shrimp, demersal fishes, mollusks, and lobster; and the coastal and littoral zones are rich in mineral resources, particularly petroleum. River discharge from several South American rivers, notably the Orinoco and Amazon, drifts through the Intra-Americas Sea and carries materials thousands of kilometers from the deltas. The large deltas are heavily impacted by anthropogenic activities, but they remain the source of rich fisheries and plankton communities. Because of the small tidal range in the Intra-Americas Sea, most deltas are wind-dominated geological features. *See* DELTA; EARTHQUAKE; MARINE FISHERIES; PLATE TECTONICS; REEF; TIDE; TSUNAMI. George A. Maul

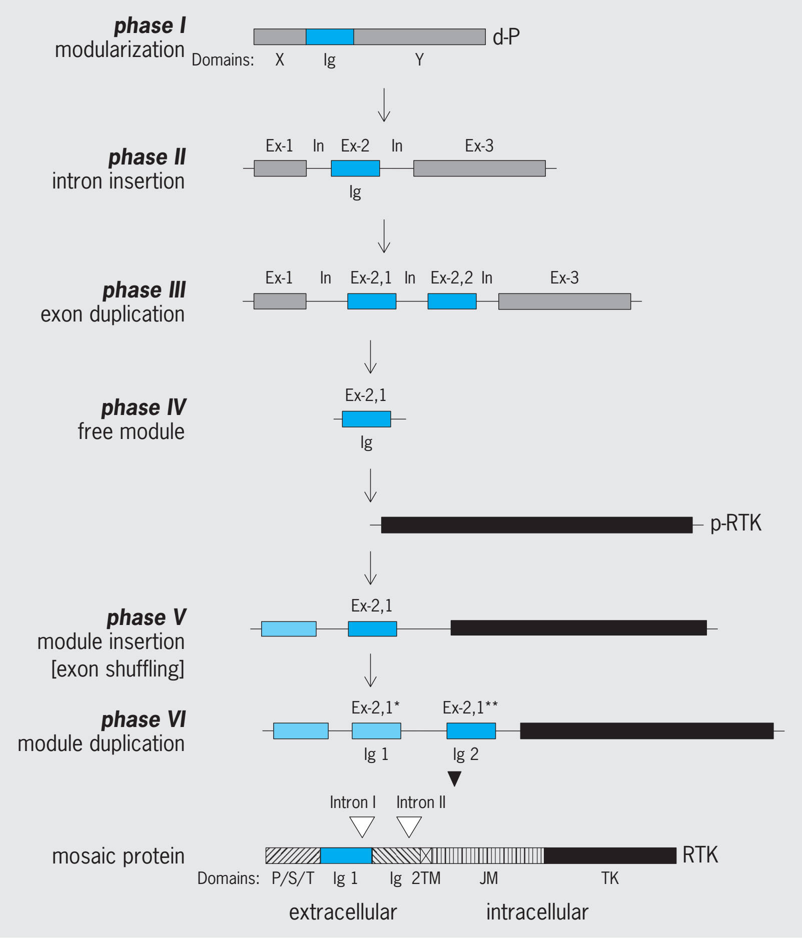
Bibliography. G. A. Maul (ed.), *Climatic Change in the Intra-Americas Sea*, United Nations Environment Programme, 1993; C. N. K. Mooers and G. A. Maul, Intra-Americas Sea Circulation, *The Sea*, vol. 2, New York, pp. 183-208, 1998.

Intron

A segment of deoxyribonucleic acid (DNA) transcribed into ribonucleic acid (RNA) as part of a longer strand of RNA, called heterogeneous nuclear RNA (hnRNA), but that does not survive the processes of messenger RNA (mRNA) maturation. Hence, genes are split into portions, termed exons, that appear in mRNAs or in structural RNAs, and introns. During the maturation of pre-mRNA to mature mRNA, a series of control processes occur that specify and narrow down the set of mRNAs that become functionally active in a given cell. These processes multiply the species of transcripts through the differential combination of exons and the inclusion of parts of introns into the functional mRNAs; thus, several mRNA products can be generated from a single gene in a controlled manner. For example, certain exons may be included or spliced out during mRNA maturation (alternative splicing of the same hnRNA molecule), or two different hnRNA molecules may form during the splicing reaction of one mRNA transcript, a process termed trans-splicing. *See* DEOXYRIBONUCLEIC ACID (DNA); EXON; RIBONUCLEIC ACID (RNA).

Split genes. Split genes were discovered in 1977 during analyses of mammalian-virus DNAs and their mRNA products. Shortly thereafter, it was shown that an intrinsic mammalian gene, for beta hemoglobin, also has protein-coding regions separated by lengths of DNA that are not translated into protein. Such discoveries were made through techniques of recombinant DNA research, and since 1978 it has become clear that most genes in eukaryotes, and a few in prokaryotes, are split. These include not just a large number of different protein-coding genes but also genes encoding transfer RNAs (tRNAs) in such diverse eukaryotes as yeast and frogs, and genes encoding structural RNAs of ribosomes in some protozoa. Introns are not limited to genes in cell nuclei but are also found in mitochondrial genes of lower eukaryotes and in some chloroplast genes.

Characteristics. The number of introns in a gene varies greatly, from one in the case of structural RNA genes to more than 50 in the collagen gene. The lengths, locations, and compositions of introns also vary greatly among genes. However, in general these sizes and locations, but not the DNA sequence, are comparable in homologous genes in different organisms. The implication is that introns became established in genes early in the evolution of eukaryotes; further, while their nucleotide sequence appears to be unimportant, their existence, positions, and sizes are significant. As an example, the genes of kinases involved in important signal transduction pathways in eukaryotes—for example, the mitogen-activated protein kinases (JNK and p38)—are interrupted by introns localized primarily in the conserved kinase subdomains. An analysis of these genes in sponges (Porifera) revealed that the positions of the introns are highly conserved from sponges to humans; moreover, in *Caenorhabditis elegans* and *Drosophila melanogaster* they are found, if present, at the same positions.



Proposed generation of mosaic proteins for sponge RTK. The gene coding for a donor protein d-P which consists of an Ig-like domain [phase I] undergoes modularization [phase II] by insertion of introns [In]. One of the created exons, Ex-2, is duplicated (forming Ex-2,1 and Ex-2,2) [phase III], and the product Ex-2,1 is liberated [phase IV] and inserted as a module into the precursor gene of RTK [p-RTK]; the process of module insertion by exon shuffling is completed [phase V]. After duplication of the two modules, the sponge RTK gene, composed of two Ig-like domains [Ex-2,1* and Ex-2.1], is completed [phase VI] and the mosaic protein is formed.**

RNA splicing. Three types or mechanisms of RNA splicing have been identified. One involves tRNA, wherein removal of a single intron from tRNA precursors is accomplished by a splicing enzyme complex that recognizes a particular RNA secondary structure, or shape. Another was discovered in studies of protozoan ribosomal RNA (rRNA), and also has been shown to be a part of the maturation of both rRNA and some mRNAs in yeast mitochondria. Excision of the intron from the precursor of *Tetrahymena* rRNA, for example, is an autocatalytic process in which the precursor RNA folds into a structure in which the intron is a loop, then cleaves itself at the ends of the intron and ligates the two parts of the mature rRNA. The reactions proceed without the involvement of protein.

Processing of RNA in mitochondria takes one of two forms. The first is the autocatalytic mechanism mentioned above, and the second involves splicing enzymes, called RNA maturases, that are actually coded for within introns of genes for other mitochondrial proteins. In an interesting form of regulation of gene expression, a maturase mRNA is translated into an enzyme that destroys its own mRNA while splicing together another mRNA. *See* MITOCHONDRIA.

The third splicing mechanism has been characterized for eukaryotic mRNAs of nuclear origin. Although little DNA sequence homology exists among introns in genes that are expressed as mRNA, there is a consistent relatedness among them at their ends (their junctions with flanking protein-coding sequences) that is involved in the removal of intron sequences from gene transcripts during RNA processing. A class of particles known as small nuclear ribonucleoproteins (SnRNPs) have been identified as additional participants in the processing of mRNAs from larger precursors. They interact with exon–intron junctions to loop out introns and facilitate RNA cleavage and exon ligation. In bacteria, certain RNAs in bacteriophage-infected cells result from the same sort of autocatalytic splicing described for protozoan rRNA.

Role. Speculation on the roles and the evolution of introns is mostly based on correlations that have been seen between domains of protein structure and the exons of genes that are defined by intervening introns. Domains in proteins are regions of a molecule that can be distinguished from others on structural and functional grounds. For example, the enzyme tyrosine kinase (TK) has several domains: among them is one that binds the substrate ATP and another that transfers the phosphate group to the protein. The presence of introns in eukaryotic structural genes and the absence in prokaryotes were explained by two different hypotheses. First, the "intron-late" theory suggests that introns arose from transposable elements which were inserted later into genes of eukaryotes. Second, the "intron-early" theory assumes that introns are the relics of the RNA world, implying that the genes-in-pieces organization of the eukaryotic genome reflects a primitive original form. The general conclusion of the intron-early theory was that the exons correspond to the building blocks from which the genes were composed by intronic recombination.

Mosaic/modular proteins and exon shuffling. Mosaic proteins or modular proteins are formed by fusion of two or more gene segments that encode different protein domains. They are assumed to be young proteins and unique to eukaryotes, especially to multicellular animals. Seeking to explain the evolution of mosaic proteins, the modularization hypothesis was formulated. It was assumed that modules suitable for exon shuffling are created by insertion of introns into the genome adjacent to exons. Members of mosaic proteins contain extracellular domains, such as receptor proteins, including the receptor protein kinases. Mosaic proteins have been identified and described from metazoan phyla including sponges. In contrast to the observations with ancient protein genes, there is strong evidence that in higher eukaryotes, in Metazoa, exon shuffling has been used to create the modular assembly of many multidomain proteins.

To test the hypothesis that the process of exon shuffling contributed to the explosive nature of metazoan radiation, studies on the gene organization of sponges, which are extant metazoans from the period prior to the Cambrian Explosion, have been performed. In one study, it was established that the sponge genes of the receptor tyrosine kinases (RTKs), enzymes found exclusively in Metazoa, contain only two introns located in the region encoding the extracellular part of the RTKs. In contrast, all metazoan genes for RTKs found in phyla younger than the Porifera contain several introns in the TK domain. This surprising fact provided experimental support for the view that ancient genes were not "in pieces." It has been suggested that introns became established in the genes of eukaryotes (and to a limited extent in bacteria) because they facilitate a genetic shuffling or rearrangement of portions of genes which encode various functional units, thus creating new genes with new combinations of properties.

The RTK gene of the sponge *Geodia cydonium* codes for a mosaic protein which is composed of several domains/modules: the Pro-Ser-Thr-rich domain, the two Ig-like domains, and the TK domain. The Ig-like domains are modules according to strict nomenclature. A schematic representation of the formation of the Ig-like module and its subsequent insertion into the RTK precursor gene is shown in the **illustration**. *See* PORIFERA.

Recombination. Genetic recombination within introns, that is, between coding units rather than within them, provides a means of genetic evolution via wholesale reassortments of functional subunits or building blocks, rather than by fortuitous recombinations of actual protein-coding DNA sequences. *See* GENE; GENETIC CODE; RECOMBINATION (GENETICS).

Werner E. G. Müller; Peter M. M. Rae

Bibliography. W. F. Doolittle, Genes in pieces: Were they ever together?, *Nature*, 272:581–582, 1978; V. Gamulin et al., Experimental indication in favor of the introns-late theory: The receptor tyrosine kinase gene from the sponge *Geodia cydonium*, *J. Mol. Evol.*, 44:242–252, 1997; W. Gilbert and M. Glynias, On the ancient nature of the introns, *Gene*, 135:137–144, 1993; B. Lewin, *Genes VII*, Oxford University Press, 2000; W. E. G. Müller (ed.), *Molecular Evolution: Towards the Origin of Metazoa* (*Progress in Molecular and Subcellular Biology 21*), Springer-Verlag, Berlin, 1998; W. E. G. Müller (ed.), *Molecular Evolution: Evidence for Monophyly of Metazoa* (*Progress in Molecular and Subcellular Biology 19*), Springer-Verlag, Berlin, 1998; W. E. G. Müller et al., Conservation of the positions of metazoan introns from sponges to humans: Evolutionary implications, *Gene*, 295:299–309, 2002; L. Patty, *Protein Evolution by Exon-Shuffling*, Springer-Verlag, New York, 1995.

Invasion ecology

The study of the establishment, spread, and ecological impact of species translocated from one region or continent to another by humans. Biological invasions have gained attention as a tool for basic research, used to study the ecology and evolution of populations and of novel biotic interactions; and as a conservation issue tied to the preservation of biodiversity. The invasion of nonindigenous (also called exotic, alien, or nonnative) species is a serious concern for those charged with managing and protecting natural as well as managed ecosystems. For example, nonindigenous species negativley affect over one-third of the U.S. Endangered Species List. *See* ECOLOGY; POPULATION ECOLOGY.

Species. Ecologists make a distinction between introduced species, meaning any species growing outside its natural habitat including cultivated or domesticated organisms, and invasive species, meaning the subset of introduced species that establish free-living populations in the wild. The great majority of introduced species (approximately 90% as estimated from some studies) do not become invasive. While certain problem invaders, such as the zebra mussel (*Dreissena polymorpha*), exact enormous economic and ecological costs, other introduced species are generally accepted as beneficial additions, such as most major food crops.

Routes of entry. Species arrive in a new region by a variety of means, including both intentional and accidental introductions. Intentional plant introductions have been promoted primarily by the horticulture industry to satisfy the public's desire for novel landscaping. However, plants have also been introduced for agriculture, for silviculture, and for control of soil erosion. Intentional animal introductions include game species brought in for sport hunting or fishing [such as ringneck pheasant (*Phasianus colchicus*) in North America, water buffalo (*Bubalus bubalis*) in Australia, and brown trout (*Salmo trutta*) in New Zealand]. Unlike these examples, intentional introductions can also include species that are not necessarily intended to form self-sustaining populations, such as those promoted by the aquarium or pet trade. Pets may be released into the wild when their disenchanted owners look for an easy method of disposal.

Species introduced accidentally are “hitchhikers.” Shipping ballast has been a major vector, first in the form of soil carrying terrestrial invertebrates and plant seeds or rhizomes, and more recently in the form of ballast water carrying planktonic larvae from foreign ports. While many species are introduced in ballast or by similar means (such as packing material before the use of styrofoam), hitchhikers can also be explicitly unwanted parasites that bypass importation and quarantine precautions. For example, many nonindigenous agricultural weeds have been imported in contaminated seed lots.

With the burgeoning of international travel and the rapid development of transportation technology, there are many new routes that species may take. In the past, global patterns of established nonindigenous species reflected the dominance of particular trade routes, for example directional movement from Europe to North America. However, trade and travel now occur between almost any two points on the planet, opening up entirely new patterns of species interchange.

Predicting invaders. Given that most introduced species will not become invaders, an important question concerns the extent to which predictions can be made as to which species will invade. The fact that the same species appear over and over invading different regions [such as the black rat (*Rattus rattus*) and water hyacinth (*Eichornia crassipes*)] is evidence that invasions are not simply random events; rather, there is such a thing as a “good invader.” Extrapolating from groups of successful invaders, biologists have compiled lists of traits that seem to confer invasiveness such as rapid growth, high reproductive rates, and tolerance to a broad range of conditions. The flaw with this approach is that saying most invaders grow rapidly is not equivalent to saying most species that grow rapidly will be invasive. To assert the latter, one has to compare the probability of invasion success for species that grow rapidly compared to species that grow slowly. This is a much more difficult task because it requires knowledge of how many species were introduced but were not successful, information that is particularly elusive for accidental introductions. However, being able to predict which species will invade has critical implications for policy regulating the importation of nonindigenous species.

Sites of invasions. Certain types of habitats seem to have higher numbers of established nonindigenous species than others. The characteristics that make a site open to invasion must be determined. For example, islands are notably vulnerable to invasions. Islands usually have fewer resident species to begin with, leading to the conjecture that simpler systems have less biotic resistance to invaders. That is, an introduced species is less likely to be met by a resident competitor, predator, or pathogen capable of excluding it. The idea of biotic resistance is also consistent with the idea that complexity confers stability in natural systems. *See* INVASION ECOLOGY.

A second generalization about invasibility is that ecosystems with high levels of anthropogenic disturbance, such as agricultural fields or roadsides, also seem to be more invaded. Increased turnover of open space in these sites could provide more opportunities for the establishment of new species. An alternative explanation is that many species that adapted to anthropogenic habitats in Europe simply tagged along as humans re-created those habitats in new places. Those species would naturally have an advantage over native species at exploiting human disturbances. A final suggestion by proponents of ecosystem management is that disturbance (including, in this context, a disruption of natural disturbance regimes, for example, fire suppression) weakens the inherent resistance of ecosystems and promotes invasion.

Spread. Early ecological models represented the spread of an invading species across a landscape with a summation of two terms: one describing growth in numbers, and the other describing movement in space, for example:

$$\begin{pmatrix}\text{Rate of}\\ \text{change in}\\ \text{population}\\ \text{density}\end{pmatrix} = \begin{pmatrix}\text{rate of}\\ \text{population}\\ \text{growth due to}\\ \text{reproduction}\end{pmatrix} + \begin{pmatrix}\text{rate of}\\ \text{dispersal}\end{pmatrix}$$

or

$$\frac{\partial N}{\partial t} = rN\left(1 - \frac{N}{K}\right) + D\frac{\partial^2 N}{\partial x^2}$$

where N is population density, t is time, r is the intrinsic rate of population increase, K is the carrying capacity, D is movement by diffusion, and x is position from the origin or release point. For such simple models, mathematicians have shown that the speed of an invasion measured in terms of a new area occupied is given by

$$\begin{pmatrix}\text{Kilometers/}\\ \text{year}\end{pmatrix} \propto \begin{pmatrix}\text{rate of}\\ \text{population}\\ \text{growth at}\\ \text{low density}\end{pmatrix} \times \begin{pmatrix}\text{rate of}\\ \text{dispersal}\end{pmatrix}$$

or

$$C = 2\sqrt{rD}$$

and will be constant after an invasion has been allowed to proceed for a sufficiently long time. The prediction of a constant rate of spread has been compared to empirical data on the movement of invading species over time, and it successfully describes the range expansion of a surprisingly diverse range of organisms, including the collared dove (*Streptopelia decaocto*) and the small white cabbage butterfly (*Pieris rapae*).

A consistent result derived from the theoretical work on range expansion is that long-distance dispersal events can have an enormous effect on the rate of spread. Moreover, the growth of newly established satellite populations can dominate the overall increase of a species in numbers or area. The obvious implication for management is that small, seemingly insignificant patches or outlying populations of a noxious weed or pest should be the focus of control efforts.

Impact. Invasive species can have several different types of impacts. First, they can affect the traits and behavior of resident organisms (for example, causing a shift in diet, size, or shape of the native species they encounter). Second, impacts can occur at the level of the population, either by changing the abundance of a native population or by changing its genetic composition. Hybridization between an invader and a closely related native can result in introgression and genetic pollution. As in the case of several endangered fish species, the endpoint can be the de facto extinction of the native species when the unique aspects of its genome are overwhelmed. Third, impacts can occur at the level of ecological communities. When individual populations are reduced or even driven extinct by competition or predation by an invasive species, the result is a decrease in the overall biodiversity of the invaded site. Finally, invaders can impact not only other species but the physical characteristics of an ecosystem as well. Shrubs that fix nitrogen with the help of symbiotic bacteria can significantly increase the nutrient pools in soil, while some deep-rooted species can dry down the water table such that native species can no longer survive.

There are two main contributing factors in determining which species have the biggest impacts: abundance and special characteristics. Invaders that reach extremely high density, such as the zebra mussel at up to 4500 per square meter, simply overwhelm all other organisms. Other species have special traits that result in an impact out of proportion to their numbers. For example, the cane toad (*Bufo marinus*) is a noxious pest in Australia in part because its toxic skin can be lethal to potential native predators.

Control. Because of the economic and conservation importance of nonindigenous species, much of invasion ecology focuses on the prevention, eradication, and control of invaders, and the restoration of sites after control. Research has emphasized the importance of early detection and eradication of problem species because of the sheer difficulty of manual and chemical control methods once populations have reached a large size.

Biological control has been touted as an environmentally friendly alternative to herbicides and pesticides. One type of biological control uses native species of insects or microbes as biopesticides to turn back invasive host populations. A second type, termed classical biological control, involves the introduction of new nonindigenous species that attack the invader in its home range. Therefore, classical biological control is also accompanied by the same risks as any intentional introduction. The uncertainties due to the current poor predictive ability of the science of invasion ecology have produced an active debate on the wisdom of using introductions to stem previous introductions. *See* ALLELOPATHY; ECOLOGICAL COMMUNITIES; ECOLOGICAL SUCCESSION; SPECIATION; SPECIES CONCEPT.

Ingrid M. Parker

Bibliography. R. Hengeveld, *Dynamics of Biological Invasions*, Chapman and Hall, 1989; H. A. Mooney and J. A. Drake, *Ecology of Biological Invasions of North America and Hawaii*, Springer-Verlag, 1986; J. L. Ruesink et al., Guilty until proven innocent: Reducing the risks of non-indigenous species introductions, *Bioscience*, 45(7):465–477, 1995; M. Williamson, *Biological Invasions*, Chapman and Hall, 1996.

Inventory control

The process of managing the timing and the quantities of goods to be ordered and stocked, so that demands can be met satisfactorily and economically. Inventories are accumulated commodities waiting to be used to meet anticipated demands. Inventory control policies are decision rules that focus on the tradeoff between the costs and benefits of alternative solutions to questions of when and how much to order for each different type of item.

Benefits of carrying inventories. For a firm, some of the possible reasons for carrying inventories are: uncertainty about the size of future demands; uncertainty about the duration of lead time for receiving deliveries; provision for greater assurance of continuing production, using work-in-process inventories as a hedge against the failure of some of the machines feeding other machines; and speculation on future prices of commodities. Some of the other important benefits of carrying inventories are reduction of ordering costs and production setup costs (these costs are less frequently incurred as the size of the orders are made larger which in turn creates higher inventories); price discounts for ordering large quantities; shipping economies; and maintenance of stable production rates and work-force levels which otherwise could fluctuate excessively due to variations in seasonal demand.

Holding costs. The benefits of carrying inventories have to be compared with the costs of holding them. Holding costs include the following elements: cost of capital for money tied up in inventories; cost of owning or renting the warehouse or other storage spaces; materials handling equipment and labor costs; costs of potential obsolescence, pilferage, and deterioration (these also involve the costs of insurance, security, and protection from natural causes such as humidity or extreme temperatures); property taxes levied on inventories; and cost of installing and operating an inventory control policy.

Pareto analysis. Inventories, when listed with respect to their annual costs, tend to exhibit a similarity to Pareto's law and distribution. A small percentage of the product lines may account for a very large share of the total inventory budget (they are called class A items, or sometimes the vital few). Aside from the class A items, and in the opposite direction, there exists a large percentage of product lines which tend to constitute a much smaller portion of the budget (they are called class C items). The remaining 20 to 30% of the items in the middle are called class B items.

The ABC analysis may help management direct more of its attention to important issues and may lead to greater cost-effectiveness. For example, if the inventory levels are checked at fixed time intervals, then the status of type A items may be reported weekly, type B items biweekly, and type C items monthly. *See* PARETO'S LAW.

Mathematical inventory theory. Mathematical inventory theory dates back to 1915 to the work of F. Harris of the Westinghouse Corporation, who derived the simple lot size formula. For independent items, economic order quantity (EOQ) means the replenishment quantity (Q) which has to be ordered to minimize the sum of the costs of holding (h dollars per item per period) and setups (and/or ordering costs: K dollars per order). If quantity discounts are not available, and if the demand (d items per period) is known and continues at the same rate through time, then the average cost per period, $Kd/Q + hQ/2$, is minimized by setting $Q = \sqrt{2Kd/h}$. This basic model was later extended to allow for various other conditions. For example, if the replenishment of the lot cannot take place instantaneously, but rather happens at a speed of r items per period, then the optimal value of Q becomes $\sqrt{2Kd/[h(1 - d/r)]}$. Other extensions of the EOQ model have been made to cover cases such as: quantity discounts in purchase prices; allowance for back orders at a cost; limited resources or facilities shared by otherwise independent items; inflation; time-varying deterministic demands and production capacities; and multiple echelons.

As for problems where the demands involve uncertainty, substantial research work has been carried out to develop inventory policies which minimize the expected values of various cost functions. Provided that information regarding the probability density function of demand is known, elegant mathematical solutions are available for independent items on a single echelon. There have also been some findings for certain multiitem, multiechelon problems; however, the amount of further research that needs to be carried out to find optimal solutions for complicated everyday inventory problems is substantial. Inventory control problems of everyday life involve many complications that can include various combinations of sequence-dependent setup times, multiitems, multiechelons with stochastic lead times, joint orders, dependent probabilistic demands, and situations where adequate information on probability density functions is not available. Even worse, the shape of the unknown probability density function may be time-varying. Regardless, mathematical inventory theory is valuable because by using the insight it provides to simpler problems, good heuristics for more complicated everyday problems can be designed.

Simulation. Computer simulation is also used for such purposes. By simulating inventory systems and by analyzing or comparing the performance of different decision policies, further insights can be acquired into the specific problem on hand and a more cost-and service-effective inventory control system can be developed.

Ordering policies. Continuous-review and fixed-interval are two different modes of operation of inventory control systems. The former means the records are updated every time items are withdrawn from stock. When the inventory level drops to a critical level called reorder point (s), a replenishment order is issued. Under fixed-interval policies, the status of the inventory at each point in time does not have to be known. The review is done periodically (every t periods).

Many policies for determining the quantity of replenishment use either fixed-order quantities or maximum-order levels. Under fixed-order quantities for a given product, the size of the replenishment lot is always the same (Q). Under maximum-order levels, the lot size is equal to a prespecified order level (S) minus the number of items (of that product) already in the system. Different combinations of the alternatives for timing and lot sizes yield different policies known by abbreviations such as (s, Q), (s, S), (s, t, S), and (t, S). Other variations of the form of inventory control policies include coordination of timing of replenishments to achieve joint orders, and adjustment of lot sizes to the medium of transportation.

Forecasting. Uncertainties of future demand play a major role in the cost of inventories. That is why the ability to better forecast future demand can substantially reduce the inventory expenditures of a firm. Conversely, using ineffective forecasting methods can lead to excessive shortages of needed items and to high levels of unnecessary ones.

Product design. Careful product design can also reduce inventory costs. Standardization, modularity, introduction of common components for different end products, and extension of the use of interchangeable parts and materials can lead to substantial savings in inventory costs.

MRP system. Material requirements planning (MRP) systems (which are production-inventory scheduling softwares that make use of computerized files and data-processing equipment) have received widespread application. Material requirements planning systems have not yet made use of mathematical inventory theory. They recognize the implications of dependent demands in multiechelon manufacturing (which includes lumpy production requirements). Integrating the bills of materials, the given production requirements of end products, and the inventory records file, material requirements planning systems generate a complete list of a production-inventory schedule for parts, subassemblies, and end products, taking into account the lead-time requirements. Material requirements planning has proved to be a useful tool for manufacturers, especially in assembly operations. *See* MATERIAL RESOURCE PLANNING.

Kanban and just-in-time (JIT). While material requirements planning systems were being developed in the United States, some Japanese manufacturers achieved widely acclaimed success with a different system. By producing components "just in time" to be used in the next step of the production process, and by extending this concept throughout the production line so that even the finished goods are delivered just in time to be sold, they obtained substantial reductions in inventories. One of the key factors for establishing just-in-time is altering the manufacturing process to drastically reduce the setup times and simplifying the ordering and procurement process so that ordering costs are cut down. The idea is to enable the producer to operate with small lot sizes, which get produced when the need arises (and not before).

Once just-in-time is established, an information system is used to determine the timing and quantities of production. Card signals—that is, visible records (in Japanese, *Kanban*)—are used to specify withdrawals from preceding production stages, and to order for production the number and type of items required. Because small batches of production have become economical, the production orders can be filled just in time. Advocates of *Kanban* characterize it as a pull process and criticize material requirements planning as a push system. Though *Kanban* is a simple idea and yields an adaptive–flexible production system, its appropriateness hinges on whether setup and ordering costs have been drastically reduced so as to allow small production batches.

Other systems. An example of application of computerized inventory control systems is IBM's Communication Oriented Production and Informational Control Systems (COPICS), which covers a wide scope of inventory-control-related activities, including demand forecasting, materials planning, and even simulation capabilities to test different strategies.

Taking into account the continuing decreases in the costs of computer hardware, the advances in computerized inventory control systems can be expected to receive even wider applications. *See* COMPUTER-INTEGRATED MANUFACTURING.

Ali Dogramaci

Bibliography. L. S. Aft, *Production and Inventory Control*, 1987; R. G. Brown, *Materials Management Systems*, 1977, reprint 1984; D. W. Fogarty, J. H. Blackstone, and T. R. Hoffmann, *Production and Inventory Management*, 2d ed., 1991; R. T. Lubben, *Just-in-Time Manufacturing: An Elegant Solution*, 1988; S. L. Narasimhan, D. N. McLeavey, and P. J. B. Billington, *Production Planning and Inventory Control*, 1995; R. J. Schonberger, *Japanese Manufacturing Techniques: Nine Hidden Lessons in Simplicity*, 1982; R. J. Schonberger, *World Class Manufacturing Casebook: Implementing JIT & TQC*, 1995; E. Silver and R. Petersen, *Decision Systems for Inventory Management and Production Planning*, 2d ed., 1995.

Inverse scattering theory

A theory whose objective is to determine the scattering object, or an interaction potential energy, from the knowledge of a scattered field. This is the opposite problem from direct scattering theory, where the scattering amplitude is determined from the equations of motion, including the potential. The equations of motion are usually linear (operator-valued) equations. An example of an inverse problem was given in an article by M. Kac entitled "Can One Hear the Shape of a Drum?" For some drums this question can be answered yes, for yet others no, but for many the question remains unanswered. *See* SCATTERING EXPERIMENTS (ATOMS AND MOLECULES); SCATTERING EXPERIMENTS (NUCLEI).

Inverse scattering theories can be divided into two types: (1) pure inverse problems, when the data consist of complete, noise-free information of the scattering amplitude; and (2) applied inverse problems, when incomplete data which are corrupted by noise are given. Many different applied inverse problems can be obtained from any pure inverse problem by using different band-limiting procedures and different noise spectra.

The difficulty of determining the exact object which produced a scattering amplitude is evident. It is often a priori information about the scatterer that makes the inversion possible.

Pure inverse scattering theory. Much of the basic knowledge of systems of atoms, molecules, and nuclear particles is obtained from inverse scattering studies using beams of different particles as probes. These probes include beams of charged particles utilizing electromagnetic interactions, weakly interacting particle beams such as neutrinos and antineutrinos, and beams of strongly interacting particles such as protons, neutrons, pions, kaons, and others.

For the Schrödinger equation with spherical symmetry or in one dimension, there is an exact solution of the inverse problem by I. M. Gel'fand and B. M. Levitan. Given complete spectral data, including (1) the energy E_i, and a normalization constant C_i for each bound state, $i = 1, 2, \ldots, N$, and (2) a weight function $w(k)$ for the continuous states, a formula exists to calculate a function $H(r, t)$. From this function the Gel'fand-Levitan equation (1)

$$K(r,t) = -H(r,t) - \int_0^r K(r,s)H(t,s)\,ds \qquad (1)$$

can be solved for K. Then the potential V sought can be calculated from Eq. (2). Furthermore, given

$$V(r) = 2\frac{dK(r,r)}{dr} \qquad (2)$$

suitable discrete and continuous approximate solutions ϕ_0, the exact solutions are given by Eq. (3)

$$\psi(r,k) = \phi_0(r,k) + \int_0^r K(r,t)\phi_0(t,k)\,dt \qquad (3)$$

where k is the wave vector. This completely solves these inverse problems. *See* ELEMENTARY PARTICLE; INTEGRAL EQUATION; NONRELATIVISTIC QUANTUM THEORY; QUANTUM FIELD THEORY; QUANTUM MECHANICS.

An application of inverse scattering to one-dimensional nonlinear evolution equations was given by C. S. Gardener and colleagues. They studied one-dimensional reflectionless potentials $V(x) = V(x, 0)$ in a time-independent linear Schrödinger equation with one, two, and so forth bound states. The time independence assures that the spectrum remains invariant under translations in time, so that the corresponding flow is isospectral. Then a second linear operator was used to generate translations in time. At large spatial distances x, the regular solution to the second time-dependent equation gave a time-dependent reflection coefficient for scattering from the potential. Inverse scattering theory was then used to convert this reflection coefficient to a time-dependent potential $V(x, t)$. The identification $V(x, t) = \phi(x, t)$ was observed to provide a solution to the nonlinear, dispersive Kortweg–de Vries equation (4), where $\phi_t = \partial\phi/\partial t$, $\phi_{xxx} = \partial^3\phi/\partial x^3$, and

$$\phi_t + 6\phi\phi_x + \phi_{xxx} = 0 \tag{4}$$

so forth. Each bound state of the time-independent Schrödinger equation corresponds to the velocity of one soliton, a localized finite-energy, solitary wave with a super stability. In ordinary stability, a solitary wave would retain its identity after a small enough perturbation; two solitons retain their size, shape, and speed with only a shift in phase even after propagating through one another. This, their superstability, requires a computer calculation to test. A number of other nonlinear equations in one space dimension and one time dimension have soliton solutions, and they are experimentally observed in plasmas, quantum optics, fluids, superconductors, and many other low-dimensional systems. *See* SOLITON.

Applied inverse scattering theory. A number of high-technology areas (nondestructive evaluation, medical diagnostics including acoustic and ultrasonic imaging, x-ray absorption and nuclear magnetic resonance tomography, radar scattering and geophysical exploration) use inverse scattering theory. Several classical waves including acoustic, electromagnetic, ultrasonic, x-rays, and others are used, extending the ability to "see" beyond visible light and beyond electromagnetic waves. For example, acoustic waves readily propagate through a thick metal container while scattering from imperfections in the wall. *See* BIOACOUSTICS, ANIMAL; COMPUTERIZED TOMOGRAPHY; MEDICAL ULTRASONIC TOMOGRAPHY; NONDESTRUCTIVE EVALUATION.

All of the inverse scattering technologies require the solution to ill-posed or improperly posed problems. A model equation is well posed if it has a unique solution which depends continuously on the initial data. It is ill posed otherwise. The ill-posed problems which are amenable to analysis, called regularizable ill-posed problems, are those which depend discontinuously upon the data. This destroys uniqueness, although solutions (in fact, many solutions) exist.

Direct scattering problems are well posed and involve solutions for a scattering amplitude function T from an integral equation of the form of Eq. (5),

$$T = V + VG_0^+T \tag{5}$$

where V and $G^+{}_0$ are given. The pure inverse problem requires the solution of Eq. (6), with T and G_0^+

$$V = -T - TG_0^+V \tag{6}$$

given. However, applied inverse problems involve general equations (7), where Y is the measured quantity, T is the direct scattering amplitude, n is noise, and R is a resolution function. The resolution function represents the effects of the measuring apparatus, especially the limit of the range of frequencies accepted. It is Eqs. (7) which are regularizable ill-posed problems.

$$Y = R \cdot T \tag{7a}$$

$$Y = R \cdot T + n \tag{7b}$$

The discontinuous dependence on data has as a consequence that two given input signals Y_1 and Y_2 can be as close together as desired everywhere, but can lead to solutions T_1 and T_2 which are as far apart as desired. The nonuniqueness means that these equations have (possibly infinite) families of solutions which are near the exact solution of Eq. (5) or (6). It is necessary to have prior knowledge about the scatterer to decide which of these solutions is the best (and even what "best" means here). The regularizer is a mathematical device for choosing one of this family. The solution chosen is called a quasisolution.

The regularizer both removes the discontinuity of the dependence of T on Y and selects a unique quasisolution. The cost of regularization is that the quasisolution selected must have some nonzero minimum spatial resolution. It is impossible to determine any finer details about the scatterer than this spatial resolution.

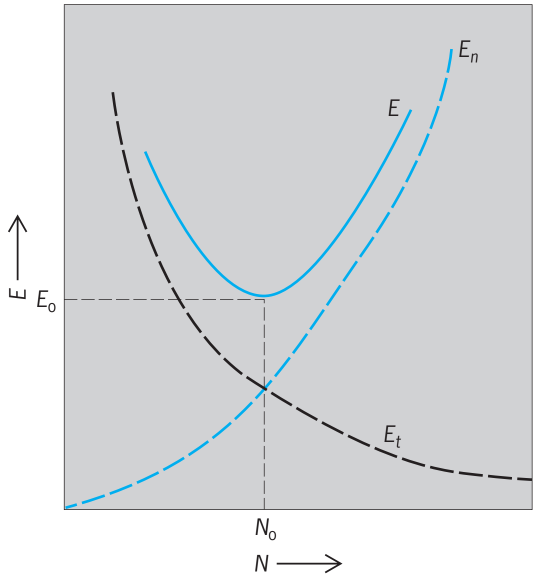
This limitation arises as follows: Let $E_t(N)$ be the truncation error, that is, the spatial resolution due to truncation of the spatial solutions at N functions of some chosen set (sines and cosines in Fourier series expansions). For a wide variety of scattering problems, it behaves as Eq. (8), where a and b are positive

$$E_t = \frac{a}{N^b} \tag{8}$$

constants which depend upon the experiment being analyzed but not N. The error due to noise E_n behaves as Eq. (9), where c and d are positive constants which

$$E_n = cN^d \tag{9}$$

depend on the experiment but not on N. For pure inverse problems, E_t is the sole source of error so that if N is large enough the error is negligible. For applied inverse problems, both errors occur so that



Example of use of Eqs. (8)–(10) to obtain the point of optimal truncation and minimum spatial resolution in an applied inverse scattering problem.

the total error is given by Eq. (10). The **illustration**

$$E = aN^{-b} + cN^{d} \qquad (10)$$

shows the truncation error E_t, the noise error E_n; and the total error E, such as that given in Eq. (10). N_0 is the point of optimal truncation. The corresponding value $E_0 = E(N_0)$ is the minimum spatial resolution.

The minimum of Eq. (10) is obtained from Eq. (11), so that N_0 is given by Eq. (12). Since $(d^2E/dN^2)_{N_0} <$

$$\left(\frac{dE}{dN}\right)_{N_0} = 0 = -abN_0^{-b-1} + cdN_0^{d-1} \qquad (11)$$

$$N_0 = \left(\frac{ab}{cd}\right)^{1/(d+b)} \qquad (12)$$

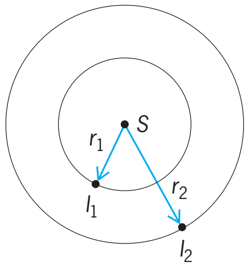
0, this is a minimum. As an example, let $a = 10^4c$ and $d = b = 2$, which gives that $N_0 = 10$ upon using Eq. (12). In this case, less than 10 terms will not yield all of the information on the scatterer available in the data; more than 10 will lose detail by allowing noise to pollute the signal.

It is the goal of applied inverse scattering theory to understand these issues and to optimize performance for each technology. Each is different, and each serves as a rich source of challenging problems. Brian DeFacio

Bibliography. M. Bertero and E. M. Pike (eds.), *Inverse Problems in Scattering and Imaging*, 1992; K. Chadan and P. C. Sabatier, *Inverse Problems in Quantum Scattering Theory*, 1977, reprint 1989; D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, 1992; G. M. Gladwell, *Inverse Problems in Scattering: An Introduction*, 1993; K. I. Hopcraft, *An Introduction to Electromagnetic Inverse Scattering*, 1992; D. Sattinger et al., *Inverse Scattering and Applications*, 1991.

Inverse-square law

Any law in which a physical quantity varies with distance from a source inversely as the square of that distance. When energy is being radiated by a point source (see **illus.**), such a law holds, provided the space between source and receiver is filled with a nondissipative, homogeneous, isotropic, unbounded medium. All unbounded waves become spherical at distances r, which are large compared with source dimensions so that the angular intensity distribution on the expanding wave surface, whose area is proportional to r^2, is fixed. Hence emerges the inverse-square law.



Point source S emitting energy of intensity I. The inverse-square law states that $I_2/I_2 = {r_2}^2/{r_2}^1$.

Similar reasoning shows that the same law applies to mechanical shear waves in elastic media and to compressional sound waves. It holds statistically for particle sources in a vacuum, such as radioactive atoms, provided there are no electromagnetic fields and no mutual interactions. The term is also used for static field laws such as the law of gravitation and Coulomb's law in electrostatics. William R. Smythe

Invertebrate embryology

The study of the development or morphogenesis and growth of the invertebrates. The same general principles of development apply to the invertebrates as to the vertebrates. Actually, much of the basic knowledge of embryology has been the result of studies on invertebrates. A common phenomenon in the invertebrates is the release of a free and independent form, the larva, before development is completed. The larvae vary considerably and are characteristic of the different animal groups.

Embryonic development begins with the formation of the gametes in a specialized cell bearing the haploid or N number of chromosomes. *See* GAMETOGENESIS.

Spermatogenesis. Number, compactness, and mobility are important for spermatozoa. Toward this end, the process of spermatogenesis consists of a stage of cell proliferation, followed by a period of progressive concentration and streamlining. The essential, heredity-determining material of the chromosomes is packed tightly into a tiny nucleus. The cytoplasm forms the locomotor apparatus, usually

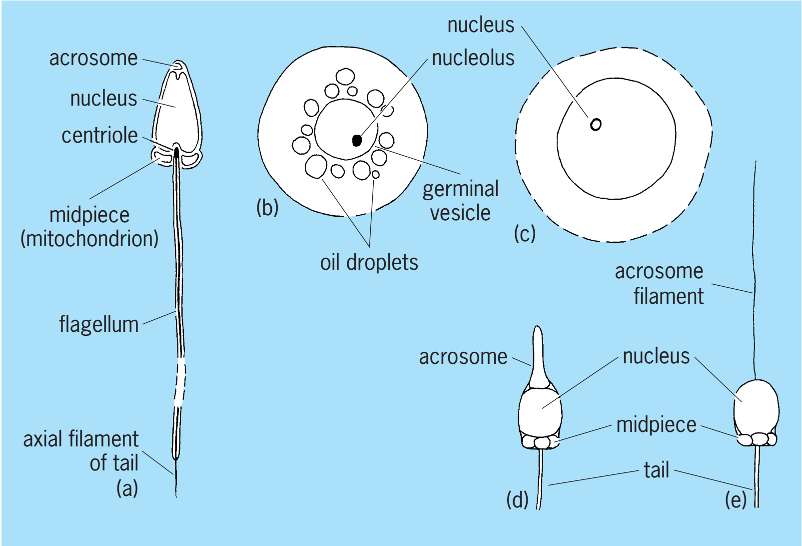


Fig. 1. Spermatozoa and fertilizable eggs. (*a*) Sea urchin spermatozoon. (*b*) Marine annelid (*Nereis*) egg, with intact germinal vesicle containing a nucleolus. Spheres surrounding germinal vesicle are oil droplets. (*c*) Marine mussel (*Mytilus*) egg, germinal vesicle broken down but polar bodies not formed. (*d*) Mussel (*Mytilus*) spermatozoa, acrosome intact. (e) *Mytilus* spermatozoa, acrosomal reaction.

a single long flagellum with a centriole at its base and a mitochondrion nearby, as well as an organelle (acrosome) for penetrating the egg coverings. Excess cytoplasm is finally discarded, and the mature spermatozoon (**Fig. 1***a*), ready to take part in fertilization, is a self-contained, stripped-down unit, carrying the hereditary message in code, and provided with enough energy source to propel it in a sustained burst of activity on its one-way trip to join an egg of its species. Millions upon millions of such cells are produced in the testis, where they remain quiescent until they are spawned. *See* SPERM CELL; SPERMATOGENESIS.

Oogenesis. The egg is designed for a very different role. It must contain enough substance to provide structural material for the formation of a set of simple organs of locomotion and digestion so that the young animal can secure food to carry on its further growth. It must also contain enough energy-yielding material to perform the work of dividing the single egg cell into body cells from which such organs can be formed and to synthesize the complex chemical substances needed to provide each of these new cells with a new nucleus. *See* OOGENESIS; OVUM.

The egg, therefore, is specialized for large size and protection of its contents, with less concern for numbers and none at all for motility. In addition, its cytoplasm possesses intrinsic capacities for differentiation and building in exact accordance with the specifications contained in its chromosomes, so that a spider egg, for example, always produces a spider and never a fly. The fact that the physical bases of these capacities have so far eluded most of the efforts directed toward their detection in no way casts doubt on their existence.

The reserve building and energy-yielding materials are stored in the egg cytoplasm as minute spheres or platelets of yolk, a stable lipoprotein substance. Eggs are large cells even without this inert material. At the end of their growth period, when they have accumulated the full amount of yolk, they are huge in comparison to the body cells of the parent animal. No invertebrate eggs, however, achieve the spectacular dimensions of bird eggs. The largest are found among the arthropods [crayfish eggs are 0.1 in. (2.5 mm) in diameter], while some marine animals have very small eggs (oyster eggs are about 65 micrometers).

During the growth period, while the egg cell is actively synthesizing yolk and increasing the amount of cytoplasm, it has a very large nucleus, the germinal vesicle (Fig. 1*b*). When it reaches full size, however, and this synthetic activity subsides, the nuclear membrane breaks down, releasing its contents into the cytoplasm. The two successive nuclear divisions of meiosis follow, but the cytoplasm, instead of dividing equally, pushes out one of the daughter nuclei each time as a polar body. These two minute bodies have no further function in development. The chromosome material left in the egg forms the egg pronucleus, which is ready to unite with the sperm pronucleus. The zygote nucleus, formed by their union, is comparable in size to those of the body cells.

Egg polarity. Many types of eggs show structural departures from radial symmetry which indicate that the unfertilized egg is organized around a bipolar axis, one end of which is called the animal pole and the other the vegetal pole. The polar bodies are given off from the animal pole, and the egg pronucleus remains in this region. When an egg contains conspicuous amounts of yolk, it is usually concentrated in the vegetal half of the egg.

Egg membranes. Since the eggs of invertebrates are often shed directly into the water of oceans and streams, or laid to develop in places where they are exposed to the drying action of air and sunlight, they are always surrounded by a protective covering. In some forms the eggs are laid in batches which may be enclosed in a leathery sac or embedded in a mass of jelly. In other cases each egg has its own separate membranous case, a layer of jelly, or a more complex system of protective structures.

Sperm and egg of each individual species have been shown by light and electron microscopy to be characteristic of its particular species. Mechanisms have evolved that are not fully understood which normally prevent the egg of one species from being fertilized by the sperm of another. Factors responsible for this species specification include egg coats, surface antigens and receptors, and environmental barriers such as substrate, temperature, and variations in salinity.

Preliminaries to fertilization. Sperm cells must complete all the nuclear and cytoplasmic changes and be fully mature before they can take part in fertilization. On the other hand, while this is true of the cytoplasm of egg cells, in most species the nuclear preparation for fertilization is incomplete when sperm entrance takes place. Moreover, the degree of incompleteness varies widely and apparently at random.

The marine annelid *Nereis* sheds its eggs with the germinal vesicle intact (Fig. 1*b*). The entrance of a spermatozoon stimulates the egg to begin the reduction divisions, and the sperm pronucleus waits within the egg cytoplasm while the egg nucleus carries out its preparation for union.

Sea urchin eggs are at the other extreme in this respect. Their reduction divisions are completed in the ovary, and the union of egg and sperm pronuclei follows immediately upon sperm entry. In many other species meiosis begins shortly before the eggs are spawned and stops at one stage or another of the process until the entrance of a spermatozoon sets it going again.

Reproduction among the invertebrates takes place in a variety of ways which differ widely from phylum to phylum. For example, following copulation in *Drosophila* the spermatozoa are stored in a special part of the female reproductive tract and released a few at a time as each ovum passes down the oviduct to be laid. The spermatozoa of squids are packaged in small bundles (spermatophores). These are placed by the male, using one of its arms which is modified for the purpose, in spermatophore receptacles on the body of the female. The eggs are fertilized as they leave the oviduct and are laid singly or in finger-shaped egg masses fixed to underwater objects.

Most of the echinoderms and many mollusks, ascidians, annelids, and coelenterates shed their eggs in tremendous numbers into the seawater where fertilization takes place. The young larvae must usually fend for themselves. In these same groups, however, some species shelter the young in seawater-containing chambers and pockets which are not actually within the body of the parent animal. This is also the case with arthropods such as crabs, which carry the larvae in masses fixed to the abdomen of the female, and some bivalves in which the young develop for a time within the mantle chamber. But whether the fertilized eggs are thus protected by the parent animal, laid in jelly masses or leathery cases or carefully constructed brood cells (bees, hunting wasps), or simply thrown into the water, each egg is an independent unit capable of developing into an adult without any further contribution from the parents.

Fertilization. The course of events of fertilization has been studied in several invertebrates, but especially in the sea urchin. Egg and sperm of the sea urchin are released into the seawater. The eggs are covered with a jelly coat to which a receptor on the plasma membrane of the fertilizing sperm binds. The plasma and outer acrosomal membranes of the sperm break down and fuse with each other as a Ca^{2+} influx occurs; the hydrolytic enzymes within the acrosome are released to lyse the egg coat. Next the inner acrosomal membrane everts by the polymerization beneath it of actin, and forms the acrosomal process which makes contact and fuses with the egg plasma membrane. The egg responds to the sperm by forming a fertilization cone. The sperm nucleus enters the egg, and its DNA swells to form the male pronucleus.

As the sperm binds to the receptors on the egg plasma membrane, the electrical potential of the egg membrane changes and establishes a rapid block to polyspermy. This prevents further sperm from making contact and fusing with the egg. With sperm-egg membrane fusion, Ca^{2+} is released to activate a series of changes in the egg. Oxygen consumption changes, a mitotic factor is released, and cortical granule release is initiated. The cortical granules are compact granules beneath the surface of the egg. At their release (**Fig. 2**), the membranes of egg and cortical granules fuse, and the contents of the granules are released into the perivitelline space surrounding the egg. One component of the granules fuses with the vitelline membrane, becomes strengthened, and lifts off the egg to become the fertilization membrane. A second component of the granules takes up Ca^{2+} and forms the hyaline layer close to the egg surface. As changes occur at the egg surface, the egg pronucleus and the sperm pronucleus with associated astral rays move toward the center of the egg, where they fuse.

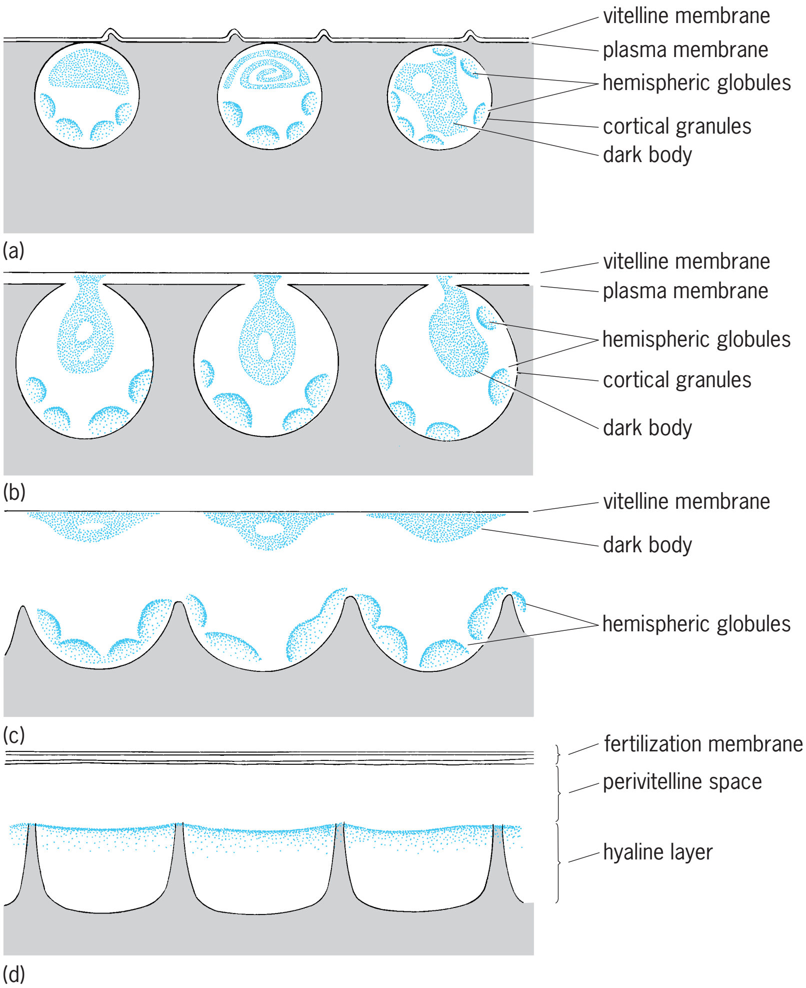


Fig. 2. Formation of the fertilization membrane in the sea urchin egg. (*a*) Surface of the unfertilized egg. (*b*) Explosion of the cortical granules: the vitelline membrane begins to be lifted up while the dark bodies are extruded, and the egg plasma membrane has become continuous with the membrane bounding the cortical granules. (*c*) The dark bodies have joined the vitelline membrane; the hemispheric globules begin to build up a layer over the new egg surface; this will then become the hyaline layer as indicated in *d*. (*d*) The dark bodies have become fused with the vitelline membrane, thus giving rise to the definitive fertilization membrane.

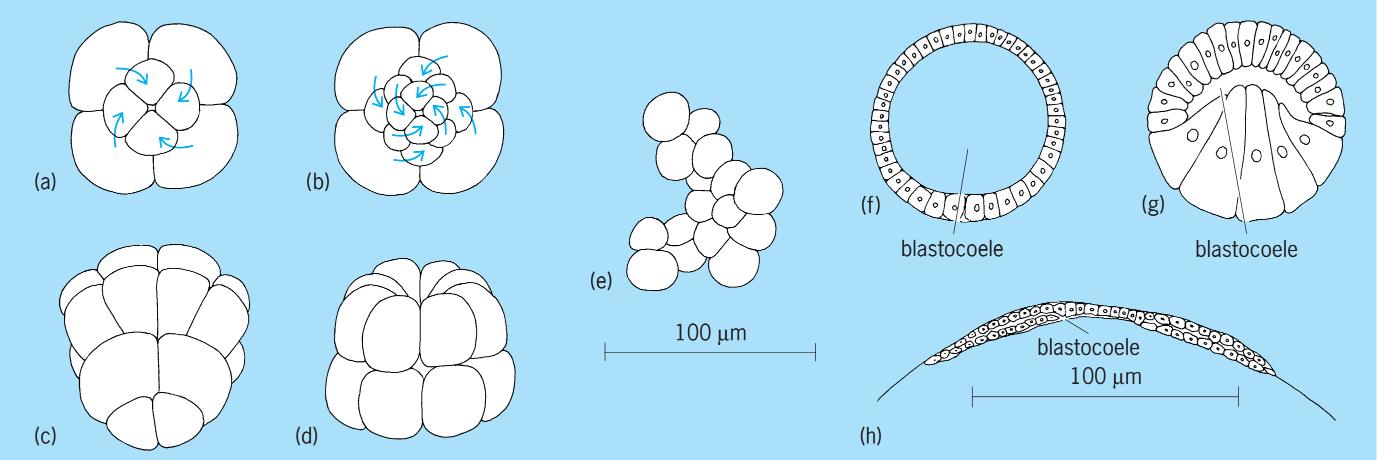


Fig. 3. Symmetry of cleavage patterns and invertebrate blastulae. (*a*, *b*) Spiral cleavage. (*c*) Bilateral cleavage. (*d*) Radial cleavage. (*e*) Irregular cleavage. (*f*) Sea urchin blastula. Relatively uniformly sized blastomeres and large blastocoele. (*g*) Annelid blastula. Blastomeres at vegetal side large, yolky; small blastocoele. (*h*) Squid blastula. Blastomeres in animal pole region only; slitlike or no blastocoele.

The union of the two pronuclei (syngamy) marks the completion of the fertilization process. The fusion forms the zygote nucleus, with the full complement of chromosomes, and the dormant egg cell has been aroused to start the series of changes which will produce a new sea urchin (**Fig. 3**).

A number of the events which have just been described are peculiar to sea urchins. With different time schedules and allowance for the individual characteristics of each species, however, these basic processes of sperm entry, aster formation, and syngamy make up the complex phenomenon of the fertilization reaction as it occurs in all animals.

Such a descriptive presentation suggests many questions about what is actually going on in terms of cellular actions and reactions. One such question, for example, concerns the nature of the contact between spermatozoon and egg surface. It is known that in many invertebrates the acrosome at the anterior tip of the sperm cell undergoes an explosive reaction at the egg surface which transforms it into a very slender filament (Fig. 1*d* and *e*). In at least some cases an enzymelike substance is also released from the acrosome which has a dissolving effect on the egg membrane. There is evidence that the penetration of the egg surface by this filament activates the egg mechanically. Whether it is also useful in drawing the spermatozoon inside the egg cytoplasm has yet to be proved.

The nature of the specificity which ensures that eggs will be entered only by spermatozoa of their own species has also been the object of a great deal of research and has not received a thoroughly satisfactory explanation. The mechanism by which an egg that has received one spermatozoon rejects all later arrivals is another problem that resists solution, but perhaps more difficult to discover than the answers to any of these questions are those concerning the cytoplasmic differences between unfertilized and fertilized eggs. *See* FERTILIZATION (ANIMAL).

Cleavage. The fertilized egg, or zygote, sets about at once to divide the huge mass of the egg into many small cells in order to restore the usual ratio between the amounts of nuclear and cytoplasmic substances. The energy for these repeated mitoses comes from the yolk, which also furnishes at least part of the materials required for synthesis of new nuclear structures. During this cleavage period, which commonly occurs during the first 12 h after fertilization, the blastomeres, as the cleavage stage cells are called, divide more or less synchronously. Generally, cleavage follows one of several patterns characteristic for large groups of animals and often correlated with the amount and mode of distribution of the yolk.

Whatever cleavage pattern is followed, the plane of the first cleavage passes through the animal pole. When the vegetal region contains a large proportion of yolk, cleavage is retarded in this area and the blastomeres tend to be larger than in the animal pole region.

Small eggs, which contain little yolk, divide completely and usually very regularly, forming a mass of cells that shows spiral (mollusks, Fig. 3*a* and *b*), bilateral (ascidians, Fig. 3*c*), or radial symmetry (echinoderms, Fig. 3*d*). Some coelenterates, however, cleave into what appear to be random masses of cells (Fig. 3*e*).

The very large eggs of squid contain a great deal of yolk concentrated at the vegetal pole. The cleavage furrows do not cut all the way through this part but restrict their activity to the living cytoplasm at the animal pole.

Insect eggs also contain a large store of yolk, which occupies the center of the elongate cells and is surrounded by a thin layer of living cytoplasm containing the egg pronucleus. Following fertilization, the nuclei alone divide and move apart in the layer of cytoplasm after each division so that they distribute themselves all around the egg. After nine such nuclear divisions have taken place (producing

512 nuclei), the cytoplasm also cleaves at the next division, forming a single layer composed of about 1000 cells surrounding the central yolk mass. *See* CLEAVAGE (DEVELOPMENTAL BIOLOGY).

Blastula stage. Among all the invertebrate forms except the insects, the result of 6–10 successive cleavage cycles is the formation of a sphere (blastula) composed of small cells which lie in a single compact layer around a central cavity (blastocoele). If the egg has contained relatively little yolk, the blastocoele is rather large (Fig. 3*f*), while it may be very small if the egg includes much yolk (Fig. 3*g*) and is little more than a slit in the squid blastula (Fig. 3*h*). *See* BLASTULATION.

Gastrula stage. The end of the brief blastula stage occurs when the process of gastrulation begins. In its simplest form, this consists in an indenting (invagination) of the blastula wall in the vegetal region (**Fig. 4***a*). Meanwhile cell division is going on steadily, and since the larva has as yet no way of taking in solid food from the outside, all the form changes which occur during this period are accomplished with the material originally present in the fertilized egg. The only addition is water (blastocoele fluid) and such dissolved substances, mostly salts, from the environment as can enter through the cell membranes. As the blastomeres become smaller and the blastular wall becomes correspondingly thinner, cells are provided to extend the vegetal indentation into a pocket (Fig. 4*b*). With the appearance of this structure (primitive digestive tract) the larva becomes two-layered, possessing an outer layer, the ectoderm, which will later produce the nervous system as well as the outermost body covering, and an inner layer, the endoderm, from which will be formed the lining of the functional digestive tract and its associated organs and glands. As the primitive digestive tract extends into the blastocoele, its opening to the outside becomes smaller and is known as the blastopore.

A modification of this process of endoderm formation occurs among some species having large, yolk-filled vegetal blastomeres (Fig. 4*c*). The small, actively dividing cells of the animal pole region spread down to cover these more inert blastomeres (Fig. 4*d*), which become the endoderm and later form the digestive organs, while the overlying ectoderm leaves a small opening in the vegetal region which corresponds to the blastopore (Fig. 4*e*). *See* GASTRULATION.

Mesoderm formation. At this time the first few cells belonging to a third body layer, the mesoderm, make their appearance by slipping from the ectoderm layer into the blastocoele. These early mesoderm cells are of a primitive sort (mesenchyme), possessing pseudopodia and often moving about freely between the ectoderm and endoderm. In sponges and coelenterates, no more highly organized middle layer is formed even in adult animals, but in the other phyla the “true” mesoderm is endodermal in origin, either being formed by successive divisions of a cell which originally belonged to the endoderm (Fig. 4*f* and *g*), as in annelids and mollusks, or separating off from the primitive digestive tract, as in *Branchiostoma* (*Amphioxus*) [Fig. 4*h* and *i*].

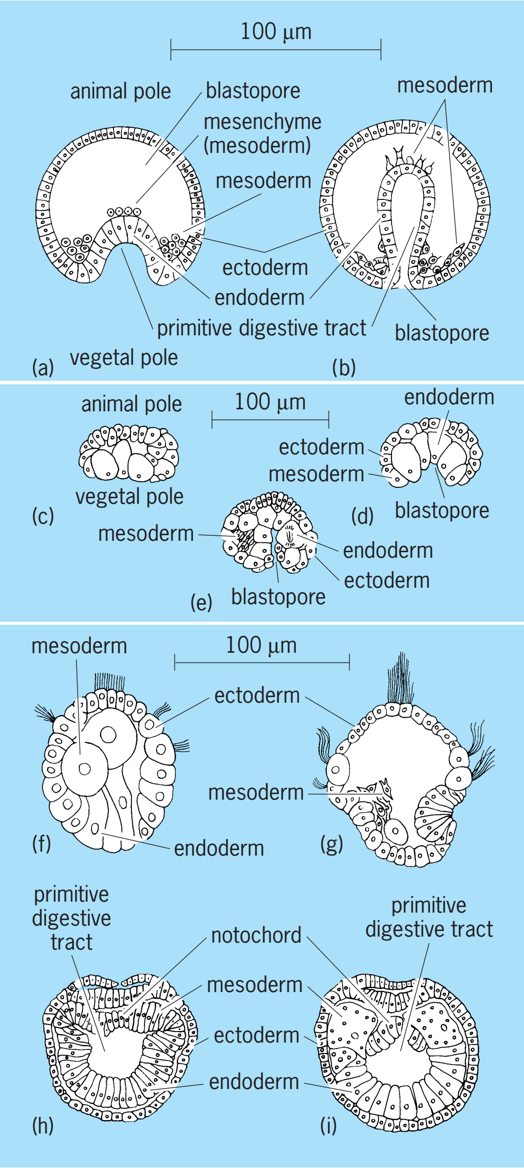


Fig. 4. Gastrulation and larval mesoderm formation. (*a*) Early and (*b*) later stage in nonyolky eggs. (*c*) Late blastula, (*d*) early gastrula, and (*e*) late gastrula in yolky egg of the snail. (*f*) Mesoderm formation in the limpet (*Patella*) shown in sections through center of blastula and (*g*) trochophore larva. (*h*) Cross section of *Branchiostoma* embryo immediately following gastrulation and (*i*) somewhat later.

In either case this mesodermal tissue spreads out between the ectoderm and endoderm, and in all phyla more advanced than the flatworms it splits through its center into an inner and an outer layer. The cavity thus formed within the mesoderm is the true body cavity in which the various internal organs lie. The outer layer of mesoderm becomes closely applied to the inner side of the ectoderm, forming body-wall muscles and other supporting layers, while the inner layer of mesoderm surrounds the endoderm with layers of muscle. The organs of circulation, excretion, and reproduction, as well as all muscles and connective tissue, are eventually formed from this mesodermal layer which surrounds the endoderm.

Later development. So far it is possible to summarize the development of invertebrate animals as a group but beyond this point each subgroup follows its own course, and these are so widely divergent that every one must be considered separately. Meaningful generalizations are not even possible within a single class in some cases, as attested to by the various modes of development occurring among the Insecta, some of which proceed directly from egg to adult form, while others go through an elaborate series of changes. *See* INSECT PHYSIOLOGY; INSECTA.

In very many species there is a sharp break in the life history when the larva, after passing through a number of morphological phases which lead from one to the next with a steady increase in size and complexity, abruptly forms a whole new set of rudimentary adult organs which take over the vital functions. This metamorphosis represents the end of the larval period. The tiny animal which it produces is for the first time recognizable as the offspring of its parents.

For more or less arbitrary reasons, the developmental processes of certain invertebrate forms have been studied very carefully so that their life histories are fully known. A few of these will be outlined in the following sections.

Molluscan Development

The eggs of *Mytilus*, the common mussel, are fertilizable just after the germinal vesicle breaks down (Fig. 1*c*). As the first polar body is given off from the animal pole, the vegetal surface of the egg bulges out to form the so-called polar lobe (**Fig.** 5*a*–*d*). The bulge disappears shortly, but will reappear at the time of the second polar body formation.

Cleavage. When the egg cleaves, the vegetal cytoplasm is segregated into a more extreme polar lobe (Fig. 5*e* and *f*) and the cleavage furrow divides the remaining material equally between two blastomeres. The constriction forming the polar lobe disappears, returning the polar lobe material to one of the blastomeres (Fig. 5*g* and *h*). The vegetal material is again segregated at the second cleavage and again mixed with one of the four blastomeres.

It is characteristic of this type of cleavage that the mitotic spindles lie aslant in the blastomeres and, moreover, regularly change the direction of their slant by 90° at each division so that a spiral pattern of blastomeres results. Such spiral cleavage is found in the mollusks and in the flat, round, and segmented worms. In modified form it is also found in the crustaceans.

Since the animal-vegetal axis is easy to recognize in such eggs, it has been possible to record the course of cleavage very accurately and to determine the role of particular blastomeres in normally developing embryos. The four-cell stage blastomere containing the polar lobe material is designated as D, and proceeding in a clockwise direction, the others become A, B, and C (Fig. 5*i*). At the third cleavage, these divide very unequally (Fig. 5*i* and *k*) into four large vegetal macromeres (1A, 1B, 1C, and 1D) and four micromeres at the animal side (1a, 1b, 1c, and 1d). *See* CELL LINEAGE.

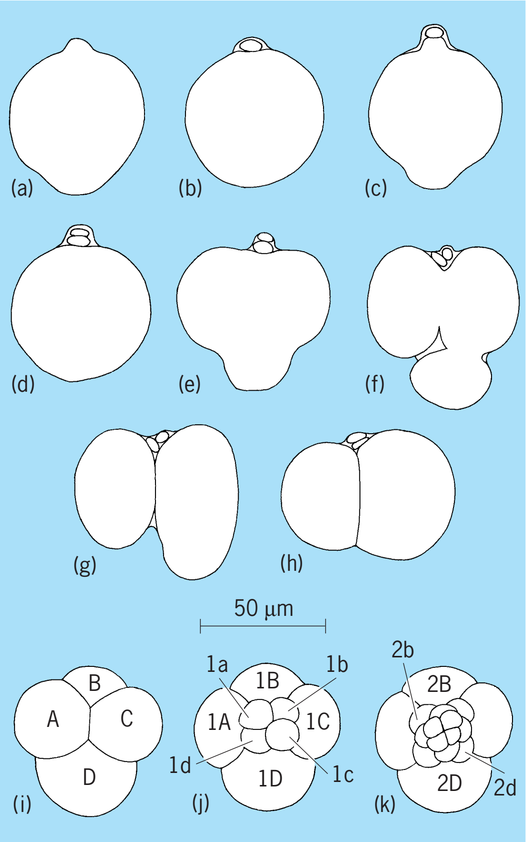


Fig. 5. Maturation and early cleavage in the mussel (*Mytilus*). (*a*) First polar body being formed at animal pole; polar lobe at vegetal pole. (*b*) First polar body completely extruded; polar lobe withdrawn. (*c*, *d*) Second polar body formation. (*e–g*) First cleavage. (*h*) Two-cell stage. (*i*) Four-cell stage. (*j*) Eight-cell stage; first quartet of micromeres (1a, 1b, 1c, and 1d). (*k*) Sixteen-cell stage with three quartets of micromeres.

Blastula stage. After two more such unequal divisions the resulting 28 micromeres have formed a hollow blastula with the four macromeres, 3A, 3B, 3C, and 3D, at its vegetal side. These then extend into the blastular cavity where their descendants will form the digestive tract, except for one of the D daughter cells produced at the next cleavage, 4d, which is set aside as the mesoderm mother cell (Fig. 4*f*).

Trochophore stage. During the succeeding cleavages some of the cells develop cilia, the blastular symmetry becomes bilateral instead of radial, and the micromeres extend down almost to the vegetal pole, thus covering the macromeres except at the small opening of the blastopore. After 24 h of development the cilia are organized into an encircling girdle and an apical tuft at the animal pole, and the larva, now called a trochophore, begins its free-swimming stage (**Fig.** 6*a*). The blastopore is shifted forward by the faster proliferation of the ectodermal cells of the other side (Fig. 6*a*) and then closed, but the larval mouth is later formed at this place. Behind

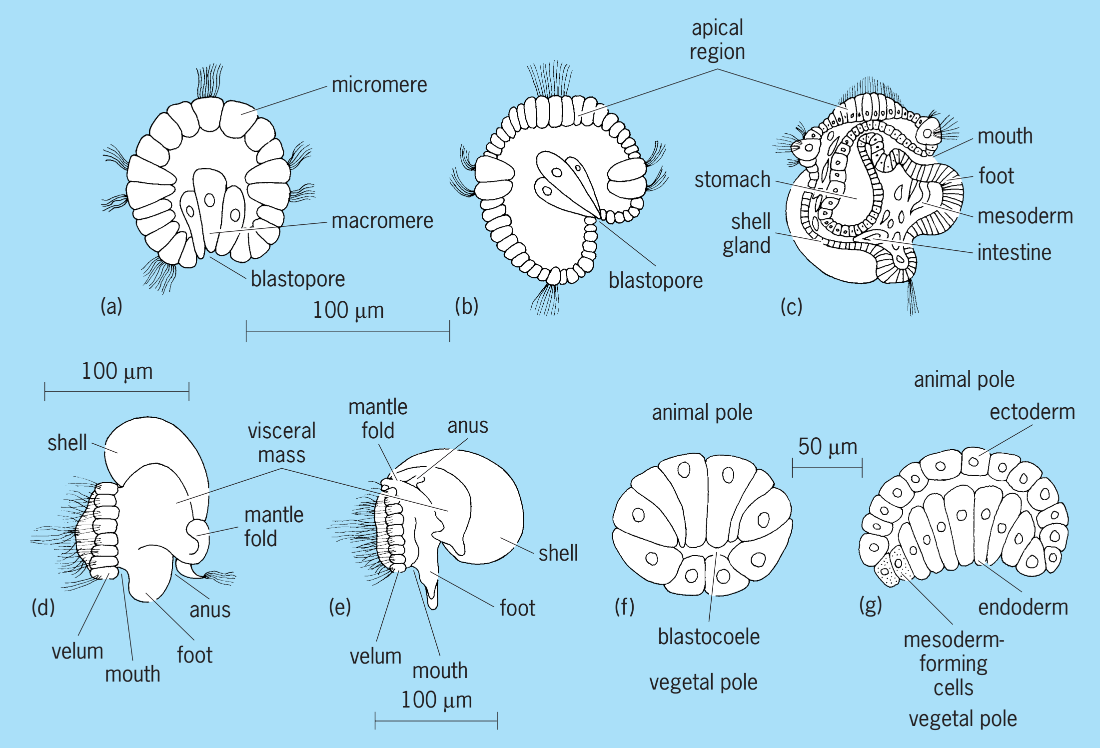


Fig. 6. Stages in the development of *Patella* and gastrulation in a tunicate. (*a*) Late blastula of *Patella*; macromeres being covered by micromeres, and blastopore at vegetal pole. (*b*) Early trochophore of *Patella*; blastopore shifted forward. (*c*) Matured trochophore larva of *Patella*. (*d*) Veliger larva of *Patella*. (*e*) Late veliger stage of *Patella*. Twisting of body has carried mantle fold and position of anus forward and turned shell about. (*f*) Late blastula of tunicate embryo. (*g*) Beginning of gastrulation in tunicate embryo.

it the endoderm forms a stomach, and a narrow tube gradually extends from this to make the intestine. The anus forms later at the place where the intestine reaches the ectoderm.

At this stage a group of ectodermal cells is forming the shell gland which will secrete the shell (Fig. 6*c*). Two small protuberances will unite and develop into the foot, and a pair of elongated pits beside the mouth will form the balancing organs. The 4d blastomere has cleaved into two cells located on either side of the mouth which are giving rise, at this stage, to two rows of mesoderm cells called the mesodermal bands.

Veliger stage. Within a week the shell gland has grown and begun to secrete the shell, and the foot is projecting prominently. The stomach increases in size and bulges into the shell cavity, and cells from the ends of the mesodermal bands form muscular attachments for the stomach and esophagus. The girdle of ciliated cells (velum) enlarges, and the rudiments of a nervous system, including eye cups, appear near the apical tuft. The larva is now called a veliger (Fig. 6*d*).

Metamorphosis. Following further development (especially of the alimentary tract, which becomes U-shaped with the mouth and anus separated from each other only by the foot), there is a sudden period of unequal growth in the two sides of the larva so that the anus is moved around to open on its neck (Fig. 6*e*). Eyes and tentacles have already been formed, and finally the young animal discards its cilium-bearing velum and takes up the adult habit of creeping about on its foot.

Sea Urchin Development

In sea urchins, the first and second cleavage planes divide the fertilized egg (Fig. 2) into four equal blastomeres, intersecting each other at right angles through the animal-vegetal axis. The third cleavage cuts through these four blastomeres horizontally. The fourth cleavage plane divides the upper four cells into eight equal-sized mesomeres, and the lower group into four very small micromeres at the vegetal pole and four large macromeres (**Fig. 7**). These 16 blastomeres are each divided about equally at the fifth cleavage, forming 32 cells, but the eight micromeres fail to divide at the sixth cleavage, giving 56 instead of 64 blastomeres at this stage. By removing certain of these blastomeres (such as the micromeres) and following later development, it has been found that the eight micromeres of 56-cell stage give rise to the first group of mesenchyme cells, the ring of eight cells just above them produces mesenchyme and endoderm, and all other blastomeres form ectoderm.

Blastulation. The blastomeres continue to divide successively, forming the hollow sphere of the blastula stage (Fig. 3*f*). By the tenth cleavage, each of the thousand or so blastomeres has developed a cilium, and the blastula has also secreted enough hatching enzyme to dissolve the fertilization membrane. After about 12 h of development at 72°F (20°C),

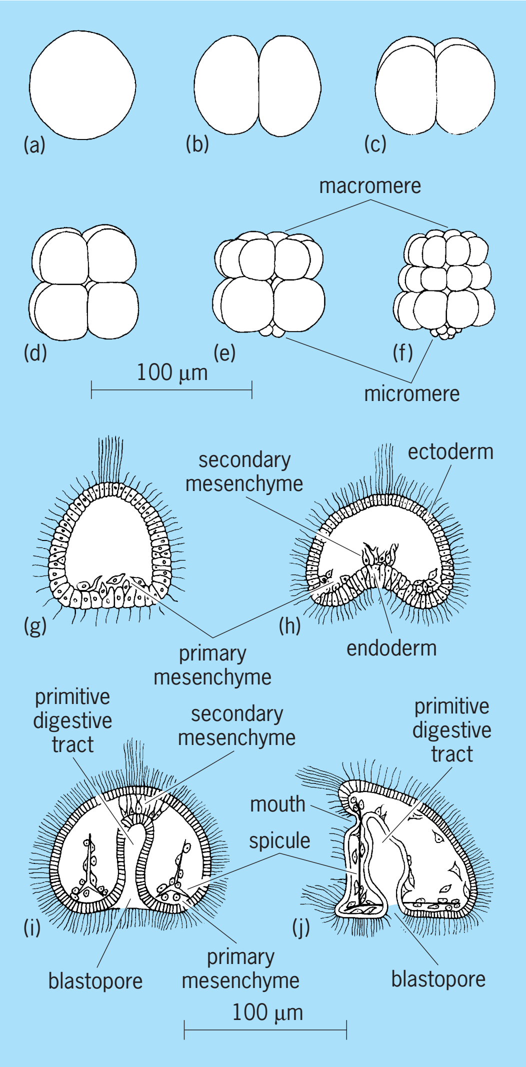


Fig. 7. Cleavage and gastrulation in the sea urchin. (*a*) Zygote. (*b*) Two-cell stage. (c) Four-cell stage. (*d*) Eight-cell stage. (e) Sixteen-cell stage. (*f*) Thirty-two–cell stage. (*g*) Vegetal region of blastula flattened; primary mesenchymal cells in blastocoele. (*h*) Beginning of invagination; secondary mesenchymal cells appearing in blastocoele. (*i*) Primitive digestive tract deepening; beginning of spicules. (*j*) Late gastrula, from side.

the larva begins its free-swimming period.

Gastrulation. Shortly afterward the larva elongates somewhat toward the animal pole, where a tuft of long, immobile cilia appears, and flattens on the vegetal side. Just before invagination begins, the cells descended from the eight micromeres slip out of the vegetal blastular wall into the blastocoele (Fig. 7*g*), forming the primary mesenchyme cells.

The gastrula stage begins about 20 h after fertilization when the center of the vegetal wall bulges inward (Fig. 7*h*). Most of the cells which form this pocket will become endoderm, but there are also some mesenchymal cells among them (secondary mesenchyme) which will develop pseudopodia that stretch across the blastocoele and, making contact with the opposite body wall, direct the inward expansion of the primitive digestive tract (Fig. 7*i*). As this deepens, the primary mesenchyme cells are building calcareous spicules with calcium taken from the seawater. These skeletal rods extend in three directions from two points beside the blastopore; as they lengthen they determine the characteristic shapes of the larval stages (Fig. 7*j*).

Pluteus stage. In the pluteus stage (**Fig. 8***a* and *b*), a mouth opening is formed where the tip of the primitive digestive tract joins the body wall. The tract begins to function, and the blastopore is changed into an anus as the larva is first able to take in food from outside.

Metamorphosis. During the month which the larva spends as a pluteus, the body increases markedly in size and two more pairs of arms are added. In preparation for metamorphosis, a structure called the echinus rudiment forms the beginnings of the adult organ systems. Within a relatively short time most of the larval body is incorporated into these organs, which are recognizable as those of a young sea urchin. Its metamorphosis is complete when it casts off the unusable parts of its larval skeleton (**Fig. 9**).

Tunicate Development

The fact that certain structures, characteristic of vertebrates and found in no other invertebrates, appear during the larval life of the tunicates forms the basis for giving these otherwise unprepossessing animals their high status at the top of the invertebrate subkingdom and makes them especially interesting from the evolutionary aspect.

Fertilization. The eggs of the tunicate *Styela* begin meiosis as they are laid, going as far as the metaphase

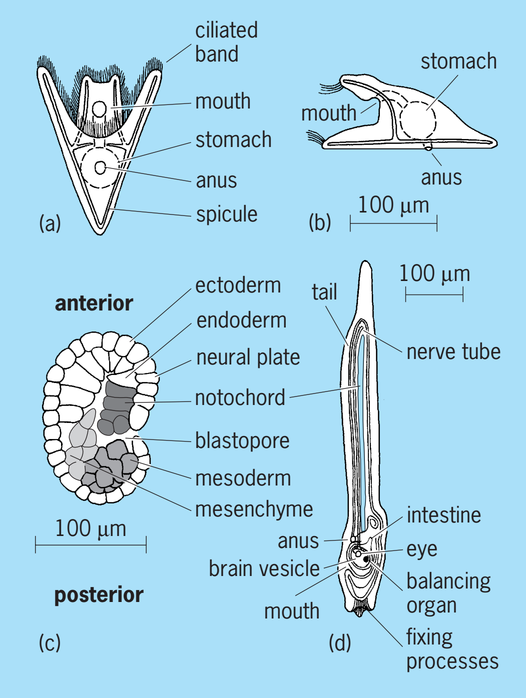


Fig. 8. Larval stages. (*a*) Ventral and (*b*) side view of pluteus larva of sea urchin. (c) Section through tunicate gastrula. (*d*) Tadpole stage of tunicate larva.

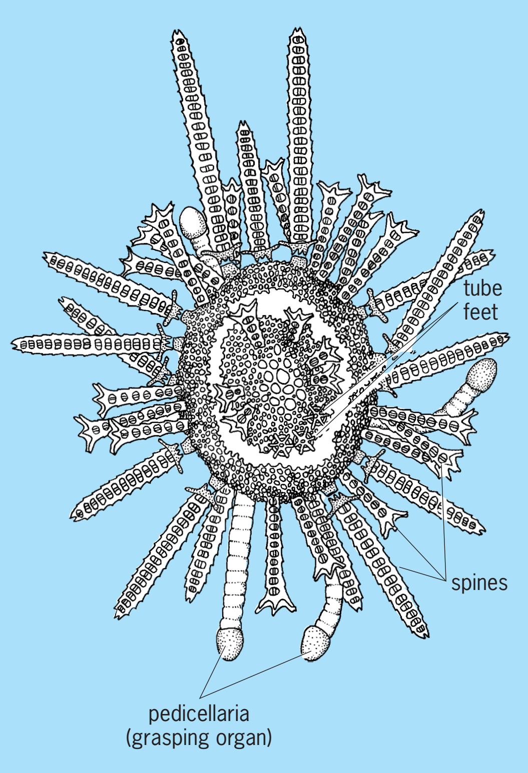


Fig. 9. Young echinoid *Peronella japonica*, after metamorphosis. (*After K. Okazaki and J. C. Dan, Metamorphosis of partial larvae of Peronella japonica Morteusen, a sand dollar, Biol. Bull., 106:83–99, 1954*)

of the first reduction division where they stop until they are fertilized. The spermatozoon penetrates the thick chorion, enters the egg at the vegetal pole, and stimulates it to proceed with meiosis. While the polar bodies are being given off, cytoplasmic streaming segregates the cell components into a yellow-pigmented region, a clear yolk-free region, and a gray yolky mass. It is possible to recognize these differently colored materials later in development and determine the role of each in body layer formation.

Cleavage. The first cleavage divides the egg into similar blastomeres. Because of the arrangement of the colored cytoplasm, its bilateral symmetry is already visible. The 16-cell stage consists of two layers of eight cells each (Fig. 3*d*), with the yellow cytoplasm contained in four of the vegetal cells. At the stage with about 40 cells a blastula is formed. The prospective ectoderm making up its animal side consists of thick columnar cells, while the future endoderm cells at the vegetal side are relatively flat (Fig. 6*f*). This difference is reversed before gastrulation begins (Fig. 6*g*).

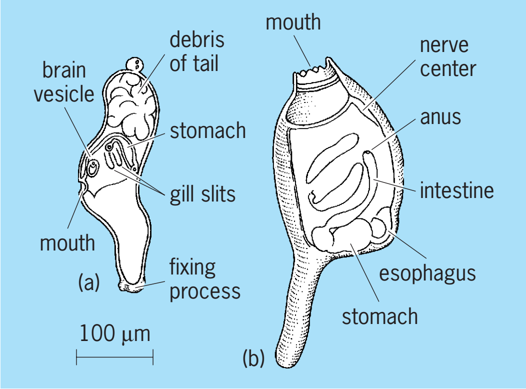


Fig. 10. Metamorphosis in tunicate. (*a*) Beginning of metamorphosis. (*b*) Young adult.

Gastrulation. The gastrula is formed by the movement into the blastocoele of the vegetal-side cells, followed by an overlapping growth of the prospective ectoderm. Within this enveloping layer the yellow cells produce mesoderm; the other vegetal cells form endoderm. As the gastrula develops, the surface layer anterior to the blastopore (Fig. 8*c*) forms neural tissue which is organized into a brain and spinal cord, while the mesoderm beneath it forms a notochord, a precursor of the vertebral column characteristic of vertebrate animals. This notochord elongates as the axis of a tail, and the larva hatches from its chorion and begins a free-swimming stage.

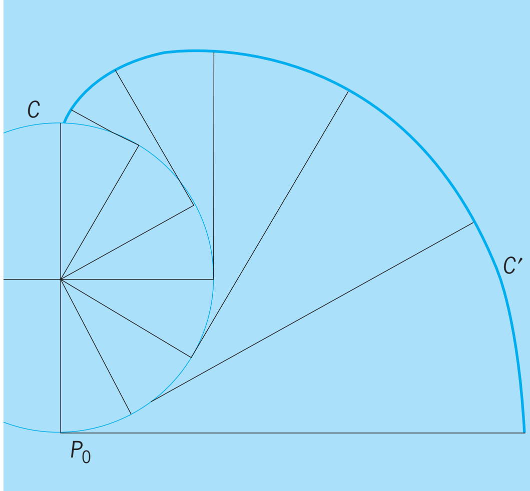
Tadpole stage. During this stage (Fig. 8*d*), the tadpole acquires an extensive but nonfunctional digestive tract, two pairs of gill slits (also characteristic of vertebrates), a "cerebral eye," and a balancing organ. At its anterior end it has three papillae with which it will fix itself to a substratum when its short tadpole stage ends.

Metamorphosis. When metamorphosis begins (**Fig. 10**), the tail ectoderm contracts strongly, bending and breaking up the notochord, nerve cord, and tail muscles which are consumed by phagocytes. The "chin" region next to the organ of fixation elongates greatly, carrying the mouth upward. A new nervous system replaces the larval one. The intestine elongates, fuses with the ectoderm to open as an anus, and forms a stomach and a liverlike gland.

A circulatory system which was started during the tadpole stage develops a muscular heart. Four new pairs of gill slits open into the pharynx. These later divide and give rise to further rows of smaller slits. The reproductive organs are formed from two masses of mesoderm cells lying near the pharynx. These develop into an ovary and a testis. Gertrude Hinsch

Involute

A term applied to a curve C' that cuts at right angles all tangents of a survey C (see **illus.**). Each curve C has infinitely many involutes and the distance between corresponding points of any two involutes is constant. If $x_i = x_i(s)$, with $i = 1, 2, 3$, are parametric equations of C, with parameter s arc length on C, all involutes C' of C have parametric equations $X_i = x_i(s) + (k - s)x_i'(s)$, with $i = 1, 2, 3$, where $x_i' = dx_i(s)/ds$, with $i = 1, 2, 3$, and k denotes an arbitrary constant. Let a length of string be coincident with a curve C, with one end fastened at a point P_0 of C. If the string is unwound, remaining taut, the other end

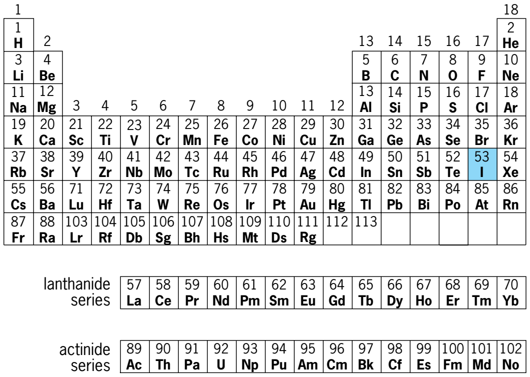


An involute C' of a curve C.

of the string traces an involute C' of C. By varying the length of the string, all involutes of C are obtained. *See* ANALYTIC GEOMETRY. Leonard M. Blumenthal

Iodine

A nonmetallic element, symbol I, atomic number 53, relative atomic mass 126.9045, the heaviest of the naturally occurring halogens. Under normal conditions iodine is a black, lustrous, volatile solid; it is named after its violet vapor. *See* HALOGEN ELEMENTS; PERIODIC TABLE.



The chemistry of iodine, like that of the other halogens, is dominated by the facility with which the atom acquires an electron to form either the iodide ion I^- or a single covalent bond —I, and by the formation, with more electronegative elements, of compounds in which the formal oxidation state of iodine is +1, +3, +5, or +7. Iodine is more electropositive than the other halogens, and its properties are modulated by: the relative weakness of covalent bonds between iodine and more electropositive elements; the large sizes of the iodine atom and iodide ion, which reduce lattice and solvation enthalpies for iodides while increasing the importance of van der Waals forces in iodine compounds; and the relative ease with which iodine is oxidized, Some properties of iodine are listed in the **table**. *See* ASTATINE; BROMINE; CHEMICAL BONDING; CHLORINE; FLUORINE.

Some important properties of iodine

Property	Value
Electronic configuration	[Kr]4d105s25p5
Relative atomic mass	126.9045
Electronegativity (Pauling scale)	2.66
Electron affinity, eV	3.13
Ionization potential, eV	10.451
Covalent radius, —I, nm	0.133
Ionic radius, I_{-}, nm	0.212
Boiling point, °C	184.35
Melting point, °C	113.5
Specific gravity (20/4)	4.940

Iodine occurs widely, although rarely in high concentration and never in elemental form. Despite the low concentration of iodine in sea water, certain species of seaweed can extract and accumulate the element. In the form of calcium iodate, iodine is found in the caliche beds in Chile. Iodine also occurs as iodide ion in some oil well brines in California, Michigan, and Japan.

The sole stable isotope of iodine is ^{127}I (53 protons, 74 neutrons). Of the 22 artificial isotopes (masses between 117 and 139), the most important is ^{131}I, with a half-life of 8 days. It is widely used in radioactive tracer work and certain radiotherapy procedures. *See* RADIOACTIVE TRACER.

Iodine exists as diatomic I_2 molecules in solid, liquid, and vapor phases, although at elevated temperatures (>200°C or 390°F) dissociation into atoms is appreciable. Short intermolecular I. . . I distances in the crystalline solid indicate strong inter-molecular van der Waals forces. Iodine is moderately soluble in nonpolar liquids, and the violet color of the solutions suggests that I_2 molecules are present, as in iodine vapor.

Although it is usually less vigorous in its reactions than the other halogens, iodine combines directly with most elements. Important exceptions are the noble gases, carbon, nitrogen, and some noble metals. The inorganic derivatives of iodine may be grouped into three classes of compounds: those with more electropositive elements, that is, iodides; those with other halogens; and those with oxygen. Organoiodine compounds fall into two categories: the iodides; and the derivatives in which iodine is in a formal positive oxidation state by virtue of bonding to another, more electronegative element. *See* GRIGNARD REACTION; HALOGENATED HYDROCARBON; HALOGENATION.

Iodine appears to be a trace element essential to animal and vegetable life. Iodide and iodate in sea water enter into the metabolic cycle of most marine flora and fauna, while in the higher mammals iodine is concentrated in the thyroid gland, being converted there to iodinated amino acids (chiefly thyroxine and iodotyrosines). They are stored in the thyroid as thyroglobulin, and thyroxine is apparently secreted by the gland. Iodine deficiency in mammals leads to goiter, a condition in which the thyroid gland becomes enlarged. *See* THYROID GLAND.

The bactericidal properties of iodine and its compounds bolster their major uses, whether for treatment of wounds or sterilization of drinking water. Also, iodine compounds are used to treat certain thyroid and heart conditions, as a dietary supplement (in the form of iodized salt), and for x-ray contrast media. *See* ANTIMICROBIAL AGENTS; ANTISEPTIC; SALT (FOOD).

Major industrial uses are in photography, where silver iodide is a constituent of fast photographic film emulsions, and in the dye industry, where iodine-containing dyes are produced for food processing and for color photography. *See* DYE; PHOTOGRAPHIC MATERIALS. Chris Adams

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; A. Varvoglis, *The Organic Chemistry of Polycoordinated Iodine*, 1992; T. Wirth (ed.), *Hypervalent Iodine Chemistry*, 2003.

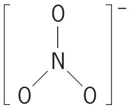
Ion

An atom or group of atoms that bears an electric charge. Positively charged ions are called cations, and negatively charged ions are called anions. When a single atom gains or loses an electron, monoatomic ions are formed. For example, reaction of the element sodium (Na) with the element chlorine (Cl), which occurs as the diatomic gas Cl_2, leads to the transfer of electrons from Na to Cl to form Na^+ cations and Cl^- anions, as shown in the reaction below.

$$2Na + Cl_2 \rightarrow 2Na^+ + 2Cl^-$$

Removal of an electron from an atom generates a species that has one more proton than electrons. For example, a sodium atom contains 11 negatively charged electrons and 11 positively charged protons. If one electron is removed, there will be 10 electrons and 11 protons, generating a positive charge on the sodium ion. Likewise, addition of an electron to an atom generates a species with more electrons than protons, which is therefore a negatively charged anion. In general, atoms of metallic elements (on the left side of the periodic table) lose electrons to form cations, while atoms of nonmetallic atoms (on the right side of the periodic table) gain electrons to form anions. Ions can bear multiple charges, as in the magnesium ion (Mg^{2+}) or the nitride ion (N^{3-}). The charge on monoatomic ions is usually the same for elements in the same column of the periodic table; for example, hydrogen (H), Na, lithium (Li), potassium (K), rubidium (Rb), and cesium (Cs) all form +1 ions. *See* PERIODIC TABLE.

Ions can also comprise more than one atom and are then called polyatomic ions. For example, the ammonium ion ($NH_4{}^+$) carries a positive charge and is composed of one nitrogen atom and four hydrogen atoms. The nitrate ion ($NO_3{}^-$) is composed of one nitrogen atom and three oxygen atoms and carries a single negative charge. Polyatomic ions are usually depicted inside brackets with superscripted charges, as shown in the structure below.



Nitrate ion

Anions and cations can combine to form solid materials called salts, which are named by the cation name followed by the anion name. For a salt composed of the polyatomic ions ammonium and nitrate, the formula is NH_4NO_3 and the name is ammonium nitrate. For monoatomic ions, the cation name is the same as the element and the anion name is the element name with the ending -ide. Thus, common table salt, NaCl, is called sodium chloride. The ratio of anions to cations must always be such that an electrically neutral material is produced. Thus, magnesium nitrate must contain one magnesium for every two nitrates, giving the formula $Mg(NO_3)_2$. *See* SALT (CHEMISTRY). H. Holden Thorp

Ion beam mixing

A process in which bombardment of a solid with a beam of energetic ions causes the intermixing of the atoms of two separate phases originally present in the near-surface region. In the well-established process of ion implantation, the ions are incident instead on a homogeneous solid, into which they are incorporated over a range of depths determined by their initial energy. In the simplest example of ion beam mixing, the solid is a composite consisting of a substrate and a thin film of a different material (**Fig. 1***a*). Ions with sufficient energy pass through the film into the substrate, and this causes mixing of the film and substrate atoms through atomic collision processes (Fig. 1*b*). If the ion dose is large enough, the original film will completely disappear (Fig. 1*c*). This process may result in the impurity doping of the substrate, in the formation of an alloy or two-phase mixture, or in the production of a stable or metastable solid phase that is different from either the film or the substrate. *See* ION IMPLANTATION.

Like ion implantation, ion beam mixing is a solid-state process that permits controlled change in the composition and properties of the near-surface region of solids. Although still being exploited experimentally and not yet employed commercially, it is expected to be useful for such applications as surface modification of metals and semiconductor device processing. Along with thin-film deposition technology, ion beam mixing should make it possible to introduce many impurity elements at concentrations that would be too high for ion implantation to be practical.

Mixing mechanisms. The ion beam mixing of a film and substrate takes place by the following

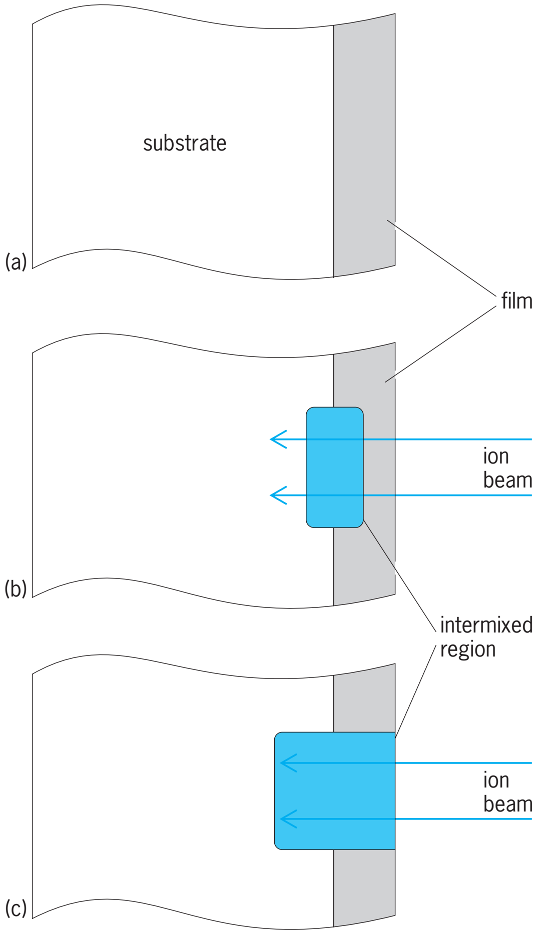


Fig. 1. Ion beam mixing of film and substrate. (*a*) Before ion bombardment. (*b*) Partial intermixing. (*c*) Complete intermixing.

mechanisms, in increasing order of complexity: recoil implantation, cascade mixing, and radiation-enhanced diffusion. In recoil implantation, an atom is driven from the film into the substrate as a result of a direct collision with an incident ion. Cascade mixing takes place when the recoil of an atom from such a collision initiates a series of secondary collisions among the film and substrate atoms. These collisions produce both forward and backward displacements and therefore lead to transfer of atoms from the substrate into the film as well as from the film into the substrate. Both recoil implantation and cascade mixing are kinematic mechanisms whose effectiveness generally increases with the mass, energy, and charge state of the incident ions and with the masses of the film and substrate atoms but is relatively insensitive to temperature. These phenomena usually yield a ratio of intermixed atoms to incident ions no greater than about 10 to 1. Since they are coincident with the atomic collisions, they occur in only about 10^{-11} s, roughly the transit time of the incident ions through the film of about 100 nm in thickness. Mixing of the film and substrate atoms by the mechanism of radiation-enhanced diffusion requires much longer times, and the rate is strongly temperature-dependent because diffusion is a thermally activated process. This mechanism is a consequence of the increase in interdiffusion coefficients that results because additional lattice defects (of which the simplest are vacancies and interstitial atoms) are formed by the atomic displacements produced by ion bombardment. Intermixing of several hundred atoms per incident ion can be achieved by heating a bombarded sample to temperatures where the defects become sufficiently mobile for rapid diffusion. *See* ION-SOLID INTERACTIONS; RADIATION DAMAGE TO MATERIALS.

Stable- and metastable-phase formation. When a sample is subjected to ion beam mixing, the nature of the solid phase obtained depends upon the composition of the intermixed region. For samples initially consisting of a thin film on a substrate, the composition obtained by complete mixing is determined by the thickness of the film. An alternate sample configuration is prepared by depositing alternating thin films of different elements on a substrate (**Fig. 2**), and passage of the energetic ions homogenizes the films on an atomic scale. For this configuration the intermixed composition can be adjusted by varying the relative thickness of the different films.

If the composition produced by ion beam mixing corresponds to a solid phase that is thermodynamically stable, this is the phase that will be obtained. Such a phase could be a solid solution of one constituent in the other, or it could be a chemical compound of the two constituents. In addition, ion beam mixing can also produce metastable phases, because it is a highly nonequilibrium process in which the kinetic energy of the bombarding ions is transferred to the target atoms and then very rapidly dissipated by atomic collision and rearrangement. For example, this technique has been used to prepare a continuous series of silver-copper solid solutions, in which the lattice constant varies almost linearly with alloy composition, and the solid solubility of nickel in silver has been increased to more than 40 times the equilibrium value. Splat cooling, the extremely rapid solidification of a liquid phase, is another nonequilibrium technique for producing metastable materials such as supersaturated solid solutions or amorphous phases. However, single-phase solid compositions can be achieved only to the extent that the constituents are miscible in the liquid phase. Thus in the silver-nickel system the effectiveness of splat cooling in increasing the solid solubility is limited

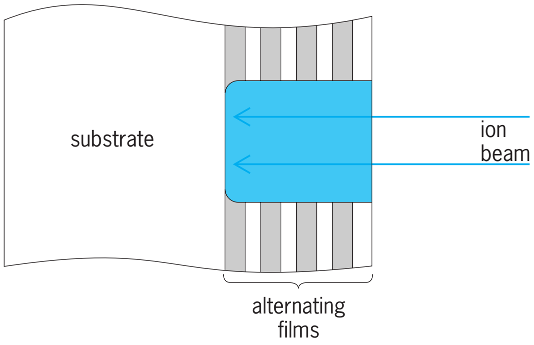


Fig. 2. Ion beam mixing of alternating films.

by the immiscibility in the liquid phase. As a solid-state process, ion beam mixing is not subject to this limitation.

Applications. While ion implantation is a well-established technique for modifying the near-surface physical and chemical properties of metals in order to improve such characteristics as wear and corrosion resistance, this technique has the disadvantage that relatively high ion doses may be necessary for effective surface modification. Thus doses of 10^{17}–10^{18} cm^{-2} are usually required to achieve impurity concentrations of 1–10 at. % in the near-surface region of about 100 nm in depth. To obtain such doses in a reasonable length of time requires high beam intensities that may be difficult to achieve for metallic elements, which are generally the ones used for direct implantation. This disadvantage can be overcome by the use of ion beam mixing, not only because the ratio of intermixed ions to bombarding ions is much greater than one, but also because the bombarding species can be ions of gases such as nitrogen and argon, for which high beam intensities are easily produced in long-life sources that are commercially available. For example, in initial demonstrations of surface modification by ion beam mixing, the wear resistance of titanium alloys has been substantially improved by the intermixing of a tin film bombarded with nitrogen ions, and the corrosion resistance of copper has been improved by the intermixing of a chromium film bombarded with argon ions.

In silicon integrated circuits, metal silicides such as titanium silicide ($TiSi_2$) and tungsten silicide (WSi_2) are widely used for low-resistance contacts and interconnects. These materials have conventionally been formed by deposition of a metal film on the silicon wafer and subsequent thermal reaction between the metal and silicon. Ion beam mixing offers the advantage of reducing the temperature required for silicide formation, since penetration of the bombarding ions through the metal-silicon interface disperses interfacial impurities (particularly native oxides of silicon), which generally form a diffusion barrier that impedes reaction. Silicides formed by ion beam mixing exhibit good surface morphology and electrical properties. *See* INTEGRATED CIRCUITS. Bor-Yeu Tsaur

Bibliography. R. E. Benenson et al., *Ion Beam Modification of Materials*, 1981; J. J. Cuomo, S. M. Rossnagel, and H. R. Kaufman (eds.), *Handbook of Ion Beam Processing Technology*, 1989; G. K. Hubler et al., *Ion Implantation and Ion Beam Processing of Materials*, 1984; F. Komarov, *Ion Beam Modification of Metals*, 1992; S. T. Picraux and W. J. Choyke, *Metastable Materials Formation by Ion Implantation*, 1982.

Ion exchange

The reversible exchange of ions of the same charge between a solution and an insoluble solid in contact with it; or between two immiscible solvents, one of which contains a soluble material with immobilized ionic groups. Ions are atoms or molecules containing charge-bearing groups. Their interactions are dominated by the electrostatic forces between charge centers. These interactions are attractive when the ions are of opposite charge, or repulsive when the ions have the same charge. Ions with a net negative charge are called anions, and those with a net positive charge are cations.

A unique property of ions is their capacity to render gases and liquids conducting, and conductivity is a universal method of detecting ions. Ions in solution are in rapid motion and have no distinct partners. Ions in an electric field migrate to the electrode of opposite charge with a velocity roughly proportional to their charge-to-size ratio. This process is known as electrophoresis, and it is one method used to separate and identify ions. *See* ELECTROPHORESIS.

Ions can also be separated on the basis of their equilibrium with a system containing immobilized ions of opposite charge. Ions can be immobilized by virtue of their location in a rigid matrix. Associated with these fixed ionic sites are mobile counterions of opposite charge. Solution ions with a higher affinity than the counterions for the fixed sites will displace them from the fixed sites and remain localized in the vicinity of the fixed sites. Simultaneously the solution is enriched in the counterions originally localized at the fixed sites. This exchange process for ions of the same charge type is called ion exchange. In a column containing the immobilized ions as part of the stationary phase and the solution of competing ions as the mobile phase, the sample ions can be separated by the repeated equilibrium steps involved as they are transported through the column until they exit it, and are detected. This is an example of ion-exchange chromatography, an important method of separating and identifying ions.

Materials

Ion exchange was first recognized over a hundred years ago, when it was found that certain clay minerals in soil can remove potassium and ammonium ions from water, with the release of an equivalent amount of calcium ions. Since that time many naturally occurring ion-exchange substances have been identified, most of them inorganic substances. Also, attention shifted from naturally occurring inorganic ion exchangers to synthetic organic polymers for most practical uses of the ion-exchange process. These materials can be prepared in large quantities with physical and chemical properties custom-made for specific applications.

Ion-exchange polymers. Phenol-formaldehyde condensation polymers have largely been replaced by materials based on styrene and divinylbenzene and, to a lesser extent, polymers prepared from divinylbenzene, or a similar cross-linking agent, and acrylic, methacrylic, or hydroxyalkyl methacrylic acids and esters. These are usually prepared in bead form by using the suspension process in which the monomers, catalysts, surfactant, and processing aids are suspended as oil droplets, which after reaction separate from the solvent as spherical beads with a narrow size distribution. By varying the experimental conditions and concentration of reagents, beads with a wide range of physical properties can be prepared. Microporous beads consist of a core of entangled polymer chains and are essentially solid beads with narrow pores possessing a small surface area. Macroporous beads are prepared by bead polymerization in the presence of a solvent that is a good solvent for the monomers but a poor swelling agent for the polymer. In this case, each polymer bead is formed from many microbeads joined together during the polymerization process to create a network of holes and channels. The macroporous beads have greater mechanical stability, permeability, and surface area than the microporous beads.

Ion exchangers prepared for the isolation or separation of cations must have negatively charged functional groups incorporated into the polymer backbone. The most common groups are sulfonic and carboxylic acids. Sulfonic acid groups are introduced by reacting the polymer beads with fuming sulfuric acid or a similar reagent. Similarly, carboxylic acid groups can be introduced by a number of common chemical reactions or by hydrolysis of the ester group or oxidation of hydroxyalkyl groups in methyl methacrylate or hydroxyalkyl methacrylate polymers, respectively. Other common functional groups used in cation exchangers include phosphoric acid and phenol and, to a lesser extent, phosphinic, arsonic, and selenonic acids. Specialty polymers containing functional groups such as 8-hydroxyquinoline or sulfoguanidine [$\text{—CH}_2\text{S(:NH)NH}_2$] are prepared for the isolation of metals by chelation.

A common approach for the preparation of anion exchangers is to react the styrene-divinylbenzene polymer with chloromethylmethyl ether in the presence of a catalyst, which adds the side chain, $\text{—CH}_2\text{Cl}$; then this chloromethylated product is treated with an amine to introduce the charged functional group. A tertiary amine produces a quaternary ammonium group, while primary and secondary amines give products that are charged only in contact with solutions of low pH. As well as simple alkyl and benzyl amines, hydroxyalkyl amines are used to introduce functional groups of the type [$\text{—CH}_2\text{N(CH}_3)_2\text{C}_2\text{H}_4\text{OH}]^+$. *See* QUATERNARY AMMONIUM SALTS.

Tentacle ion exchangers consist of an insoluble matrix, copolymerized from oligoethyleneglycol, glycidylmethacrylate, and pentaerythroldimethacrylate, to which are grafted polymerized chains of acrylamide derivatives that are approximately 15–50 units in length. An average of 18 charged groups is covalently bound to each of these tentacles. This arrangement markedly reduces the contact between the analyte and the matrix, thus suppressing nonspecific support interactions, while also allowing ionic interactions between the ion exchanger and the analyte that are sterically impossible with conventional fixed-site ion exchangers. These materials are used for the isolation of biopolymers without loss of activity.

Ion exchangers with a low exchange capacity and fast mass-transfer properties are required for ion chromatography that uses conductivity detection. The time and temperature of the chemical reactions used to introduce the ionic functional groups into the polymer are controlled such that the reactions are limited to the external surface of the beads. Surface-functionalized anion-exchange polymer beads containing quaternary ammonium ions are difficult to prepare directly. An alternative approach is the preparation of agglomerated anion-exchange beads, prepared by taking surface-sulfonated beads and coating them with a layer of quaternary ammonium functionalized latex particles. The core particles are 5–30 micrometers in diameter, and the latex particles are much smaller at about 0.6–0.9 nanometer. The negatively charged sulfonate groups attract the positively charged quaternary ammonium ions and bind the latex particles very strongly to the surface of the polymer bead. Since only a small fraction of the quaternary ions on the latex particles are involved in binding of the anion-exchange layer, the remaining groups are free to function as an anion exchanger. This type of structure creates a thin anion-exchange pellicular layer within which the ion-exchange separation takes place, and an oppositely charged underlayer that excludes anions from the core by repulsion.

Silica-based materials. These are used primarily in chromatography because of the favorable mechanical and physical properties of the silica (SiO_2) gel support matrix. Ion-exchange groups are introduced by reacting the surface silanol groups of the porous silica particles with silanizing reagents containing the desired functional group (R). This produces a siloxane-type (Si—O—Si—R) bonded phase, where the bonded layer can be monomolecular or polymeric depending on the choice of reagent and experimental conditions. Typical R groups include propylsulfonic acid [$\text{—(CH}_2)_3\text{SO}_3^-$], benzenesulfonic acid [$\text{—C}_6\text{H}_4\text{SO}_3^-$], butyric acid [$\text{—(CH}_2)_3\text{CO}_2^-$], propylbenzoic acid [$\text{—(CH}_2)_3\text{C}_6\text{H}_4\text{CO}_2^-$], aminopropane [$\text{—(CH}_2)_3\text{NH}_2$], and dimethylaminopropane [$\text{—(CH}_2)_3\text{N(CH}_3)_3^+$]. Alternatively, ion exchangers based on silica and other inorganic oxides can be prepared by polymer encapsulation. In this approach a prepolymer such as poly(butadiene-maleic acid) is deposited as a thin film over the porous support and is subsequently immobilized by radical-induced in-place cross-linking of the polymer chains. This creates an ion-exchange polymer as a thin skin covering all or a sizable part of the support surface.

Hydrous oxides. Hydrous oxides of elements of groups 14, 15, and 16 of the periodic table can be used as selective ion exchangers. The most important hydrous oxides used for the separation of organic and inorganic ions are alumina ($\text{Al}_2\text{O}_3\cdot n\text{H}_2\text{O}$), silica ($\text{SiO}_2\cdot n\text{H}_2\text{O}$), and zirconia ($\text{ZrO}_2\cdot n\text{H}_2\text{O}$). Silica, by virtue of the presence of surface silanol groups, is used as a cation exchanger at pH > 2. Alumina is amphoteric and can be used as an anion exchanger at low pH and a cation exchanger at high pH. Alumina has the advantage over silica of being chemically stable over a wide pH range. The ion-exchange capacity of silica and alumina is controlled by the pH of the solution in contact with the oxides, since this controls the number of ionized surface functional groups. Alumina is used to isolate nitrogen-containing drugs and biochemically active substances from biological fluids, thus minimizing matrix interferences in their subsequent chromatographic analysis. Group 1 and 2 cations can be separated by chromatography on silica, with aqueous lithium salts as eluants. Zirconium phosphate is not a hydrous oxide, but it has many properties in common with the hydrous oxides. It has a high selectivity for the ions rubidium (Rb^+), cesium (Cs^+), strontium (Sr^{2+}), and barium (Ba^{2+}) and is used to remove long-lived $^{90}Sr^{2+}$ and $^{137}Cs^+$ from radioactive wastes. *See* PERIODIC TABLE; PH; RADIOACTIVE WASTE MANAGEMENT.

Synthetic carbohydrate polymers. Ion-exchange polymers based on cross-linked dextrans, agarose, and cellulose are used primarily in biochemistry for the isolation and separation of polyelectrolytes such as proteins, enzymes, and nucleotides. The ion-exchange groups are introduced into the hydrophobic matrix by reacting a suitable reagent with the hydroxyl groups on the saccharide subunits. Typical functional groups include carboxymethyl, phosphoric acid, propylsulfonate, diethylaminoethyl, and quaternary amines.

Characteristic Properties

Ion exchangers are broadly categorized as belonging to one of two groups. Strong ion exchangers, such as those containing sulfonate or quaternary amine groups, remain ionized over the full range of normal operating conditions. The ionization of weak ion exchangers, such as those containing carboxyl or secondary amine groups, depends on the pH of the medium. Their degree of ionization governs their ion-exchange capacity.

Capacity. The capacity of an ion exchange is the quantity of exchangeable ions per unit quantity of exchanger. It is usually expressed in gram-equivalents of ions per kilogram of exchanger, or more conveniently, milliequivalents per gram (meq/g) on a dry basis. General-purpose polymeric ion exchangers have capacities of about 3–5 meq/g. For ion chromatography a high capacity is undesirable because of the need to separate ions with eluants of low ionic strength to facilitate their detection by conductivity, and typical exchange capacities are in the range 0.001–0.1 meq/g.

Volume. The amount of cross-linking reagent used in the preparation of bead polymers controls the extent to which the dry polymer swells when placed in solution, the speed with which ions can diffuse in and out of the ion exchanger, and to some extent the selectivity of the ion exchanger for ions of different charge-to-size ratios. Styrene-divinylbenzene polymers containing less than about 8 mole % of divinylbenzene are classified as soft. When suspended in a solvent, the dry polymer absorbs the solvent, changing its volume. The observed change in volume depends on properties of the solvent as well as those of the polymer. When the polymer comes into contact with solvent, the outermost functional groups are solvated, and the randomly arranged polymer chains unfold to accommodate the larger solvated ions. A very concentrated internal solution of fixed ions and counterions exists, and the mobile counterions tend to diffuse out of the exchanger to be replaced by solvent in an attempt to reduce the difference in ionic strength between the internal and external solutions. This generates the swelling pressure, which may be as high as several hundred atmospheres for a high-capacity exchanger. The cross-linking of the polymer chains resists this swelling with different degrees of success that depend on structural and environmental factors. High-pressure liquid chromatography requires materials that are rigid and do not swell or change volume in different solvents. A high level of cross-linking is invariably required in this case.

Size. Polymer beads are available in a wide range of sizes and in narrow size ranges. Large beads (2–0.1 mm in diameter) are used in industrial processing, beads of intermediate size (0.2–0.04 mm) for sample processing and chromatography at limited pressures, and small beads (0.04–0.004 mm) for high-performance liquid chromatography at high pressures. Silica-based materials are available in a similar size range and are widely used in high-pressure liquid chromatography. The carbohydrate-based polymers are generally unstable to high pressures and are usually available in the large and intermediate size ranges only. The superior mass-transfer and favorable flow anisotropy properties of small-diameter particles are essential for modern liquid chromatography. Diffusion of ions through polymer gels is perhaps only one-tenth as fast as in liquids, and the only practical way to reduce this contribution to band broadening in chromatography is to minimize the particle diameter to reduce the average path length for migration of ions through the beads. In the surface-functionalized polymers used in ion chromatography, the migration distance for the exchangeable ions is restricted to the thickness of the active layer at the bead surface, which is only a few micrometers thick. From the perspective of performance characteristics, the silica-based materials are usually better, but the difference in properties between the polymer bead exchangers and the silica-based materials is no longer as significant as before.

Stability and adsorption. The stability and adsorption properties of the matrix are important in selecting an ion exchanger. Silica-based materials, for example, are restricted to use within the pH range 2–8 to avoid dissolution of the silica backbone at high pH and hydrolysis of the surface-bonded siloxane ligands at lower pH. Poly(styrene-divinylbenzene) polymers can be used at all pHs. For biopolymers the choice of matrix may be critical to the recovery of materials with preservation of biological activity. *See* LIGAND.

Applications

Ion exchange has numerous applications for industry and for laboratory research. By the quantity of materials used, water conditioning is the most important. Ion exchange is one of the primary analytical methods used to identify and quantify the concentration of ions in a wide range of environmental, biological, and industrial samples.

Water softening. Natural water from rivers and wells is never pure; it is usually hard, that is, it contains calcium and magnesium salts that form curds with soap and leave hard crusts in pipes and boilers. Hard water is softened by passage through a cartridge or bed of cation exchanger in the sodium form (the mobile counterions are sodium in this case). The calcium and magnesium ions have a higher affinity for the fixed ion-exchange sites than sodium ions and replace the sodium ions on the exchanger. The sodium ions do not cause hardness, so their increased concentration in the conditioned water is not a problem. Eventually the cation exchanger becomes exhausted, that is, fully loaded with magnesium and calcium ions, and must be regenerated or replaced. The cation exchanger is regenerated by passing a concentrated solution of sodium chloride through the bed of ion exchanger. *See* WATER SOFTENING.

Deionized water. Many industrial and laboratory processes require a supply of pure water with a very low concentration of salts. This can be achieved by passing water through a bed of mixed strong cation exchanger in the hydrogen form and a strong anion exchanger in the hydroxide form. The cation exchanger removes all the cations from the water by replacing them by hydrogen ions. The anions are removed by the anion exchanger and replaced by hydroxide ions. The hydrogen and hydroxide ions combine to form water. The purity of the water can be monitored by its conductivity. Inclusion of a cartridge containing carbon to adsorb neutral organic compounds produces a reliable supply of very pure water containing only parts-per-billion levels of ions and organics. Mixed-bed ion exchangers tend to be more efficient, but they are difficult to regenerate. The series coupling of cation-exchange and anion-exchange cartridges allows for ease of regeneration (**Fig. 1**).

Environmental reclamation. Toxic ions such as mercury (Hg^{2+}), lead (Pb^{2+}), chromate ($CrO_4{}^{2-}$), and ferrocyanide ($(Fe(CN))_6{}^{4-}$ are removed by ion exchange from industrial wastewaters prior to their discharge into the environment. Ion exchangers are used to recover precious metals such as gold (Au^+), platinum (Pt^+), and silver (Ag^+) in a useful form from mine workings and metalworking factories. Ion exchange is frequently used to decontaminate waste and concentrate radioactive elements from the nuclear industry. Uranium can be recovered from low-grade ores by leaching with dilute sulfuric acid, then absorbing the uranium sulfate complex ions on a strong anion exchanger, which has a high affinity for uranium sulfate. *See* HYDROMETALLURGY; SOLUTION MINING; URANIUM METALLURGY.

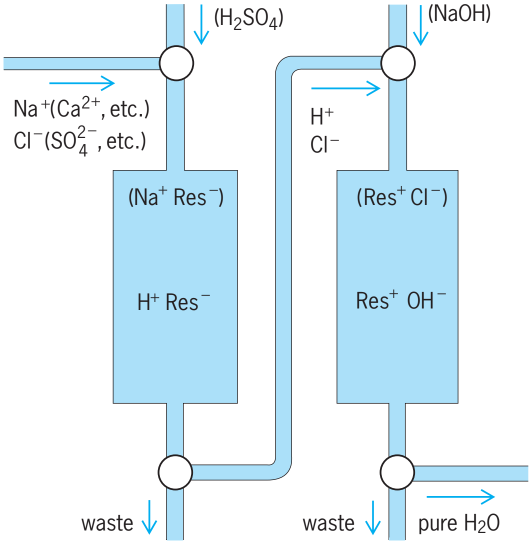


Fig. 1. Two-stage deionization by ion exchange. Valves are for regeneration with sulfuric acid (H_2SO_4) and sodium hydroxide (NaOH).

Chemical analysis. Ion exchange is used on the laboratory scale for isolation and preconcentration of ions prior to instrumental analysis and to obtain preparative scale quantities of material for use in laboratory studies. For example, organic acids can be isolated from urine with a strong cation exchanger free of interference from the bulk of nonionic water-soluble organic compounds normally present in urine. Trace quantities of ions in ice from the polar regions can be concentrated by ion exchange for subsequent separation and identification by ion chromatography. Ion exchange is often employed in conjunction with activation analysis to isolate individual elements for quantification by radiochemical detection. Isolation schemes for individual elements in complex mixtures by ion exchange employing multiple changes in buffers and complexing agents are well known. These methods are slow and tedious but reliable, and they serve as reference methods to check the validity of fast routine methods of analyzing complex materials. *See* ACTIVATION ANALYSIS; ANALYTICAL CHEMISTRY.

Biotechnology. Biotechnology requires reliable, efficient methods to purify commercial-scale quantities of proteins, peptides, and nucleic acids for use in the pharmaceutical, agricultural, and food industries. Ion exchange is widely used in the isolation and purification of these materials. Typical applications include the removal of ionic compounds used in the production process, the elimination of endotoxins and viruses, the removal of host-cell proteins and deoxyribonucleic acid (DNA), and the removal of potentially hazardous variants of the main product. Since the biological activity of proteins is closely related to their three-dimensional structure and this structure is labile, one of the main reasons for choosing ion-exchange methods for the isolation and purification of proteins is that weakly denaturing conditions can be employed throughout the process steps. For commercial exploitation of biotechnology products, the retention of biological activity is as important as recovery; and if the product is destined for human or animal consumption, biological impurities with adverse reactions must be reduced to a specified control range.

Chromatography. Modern chromatographic techniques employ ion exchangers of small particle size and favorable mass-transfer characteristics, and operate at high pressures, providing better resolution of mixtures in a shorter time than with conventional gravity-flow-controlled separations. In clinical laboratories, ion exchange has long been employed as the basis for the routine, automated separation of amino acids and other physiologically important amines used to identify metabolic disorders and to sequence the structure of biopolymers. Under typical conditions the amino acids are separated in the protonated form on a strong cation exchanger with stepwise changes in pH by using various combinations of citrate and borate buffers. The amino acids are usually detected as their fluorescent products after a postcolumn reaction with ninhydrin or *o*-phthalaldehyde. *See* AMINO ACIDS.

Probably the oldest and most common method of separating carbohydrate mixtures is by ligand-exchange chromatography using polymeric cation exchangers loaded with metal ions, for example calcium, for separating sugar and silver to determine the oligomer distribution of oligosaccharides, with aqueous organic solvents as the mobile phase. Retention in this case is governed by a combination of size exclusion and electrostatic attraction between the electronegative sugar oxygen atoms and the electropositive metal cations. Carbohydrates are also separated by anion-exchange chromatography after they are converted to negatively charged borate complexes or as their negative ion form at high pH in aqueous solution.

Ion chromatography is an ion-exchange-based separation system that changed the practice of ion analysis by replacing many tedious wet-chemical procedures with a simple automated instrument that can determine several ions simultaneously in a single method (**Fig. 2**). Its evolution into a major instrumental method of analysis was a synergistic development of efficient low-capacity ion-exchange column packings that allowed the use of low-concentration eluants containing competing ions with a high affinity for the ion exchanger, together with the provision of continuous on-line detection based on conductivity monitoring of the column effluent. Ion chromatography has been used extensively for the determination of ions in drinking water, wastewater, plating baths, detergents, foods, and many other applications. It is almost always the method of choice for determining anions and organic cations and has become popular for the separation of metal ions. *See* CHROMATOGRAPHY; LIQUID CHROMATOGRAPHY.

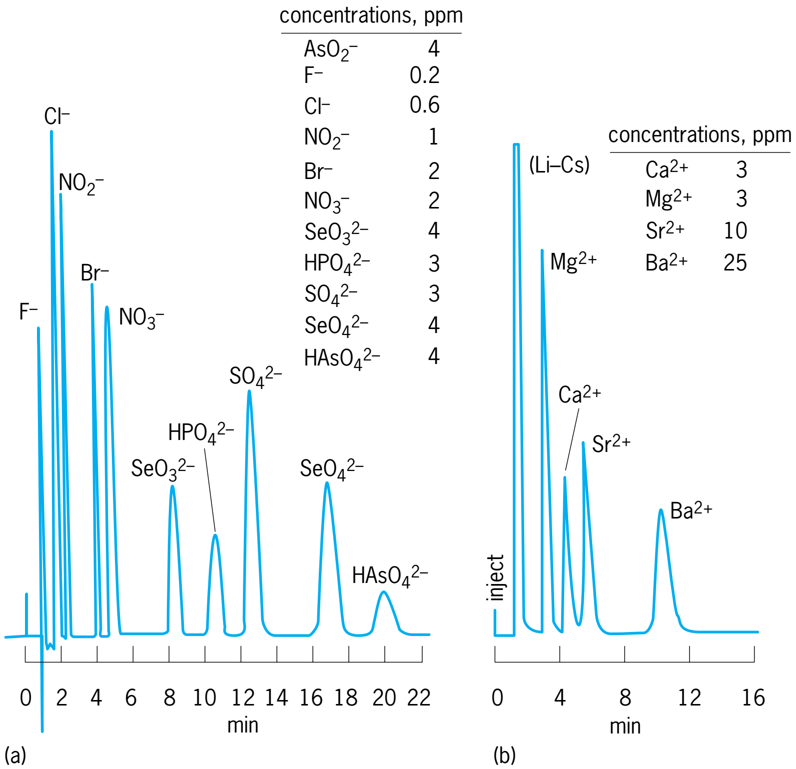


Fig. 2. Chromatograms showing separation of (*a*) common anions and (*b*) alkaline-earth cations by ion chromatography with conductivity detection. (*After C. F. Poole and S. K. Poole, Chromatography Today, Elsevier, 1991*)

Ion-pair chromatography. Ion-pair chromatography provides an alternative method of separating ions without the use of an ion-exchange column. A temporary ion-exchange layer is formed by the adsorption of hydrophobic ions onto the surface of a reversed-phase column packing material. The hydrophobic ions are added to the mobile phase in a fixed concentration in equilibrium with the ions adsorbed onto the stationary phase, and are present throughout the separation. Suitable hydrophobic ions for separating anions are tetraalkylammonium salts, and for the separation of cations, alkanesulfonate salts. The most important variables in optimizing the separation of ions by ion-pair chromatography are the concentration of the ion-pair reagent, the concentration of organic solvent in the mobile phase, and the mobile-phase pH. An advantage of ion-pair chromatography is that ions and neutral molecules can be separated at the same time by using the efficient column packings developed for reversed-phase chromatography. Colin F. Poole

Membranes

Ion-exchange membranes are a class of membranes that bear ionic groups and therefore have the ability to selectively permit the transport of ions through themselves. In biological systems, cell membranes and many other biological membranes contain ionic groups, and the conduction of ions is essential to their function. Synthetic ion-exchange membranes are an important field of technology. They are used in fuel cells, electrochemical processes for chlorine manufacture and desalination, membrane electrodes, and separation processes.

Ion-exchange membranes typically consist of a thin-film phase, usually polymeric, to which have

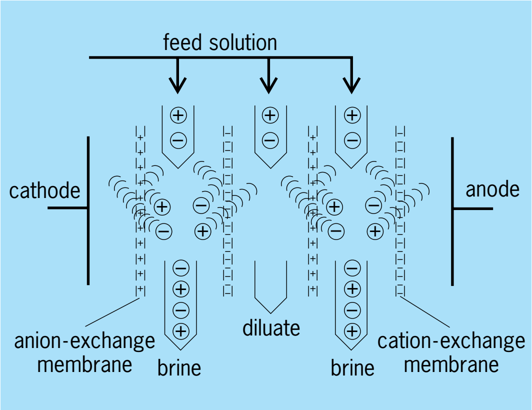


Fig. 3. Electrodialysis process, in which the ions of the saline water are separated under the influence of an electric field. (*After H. Strathmann, The principle of electrodialysis, in P. M. Bungay et al., eds., Synthetic Membranes: Science, Engineering and Applications, D. Reidel Publishing Co., 1986*)

been attached ionizable groups. Numerous polymers have been used, including polystyrene, polyethylene, polysulfone, and fluorinated polymers. Ionic groups attached to the polymer include sulfonate (—SO_3^-), carboxylate (—COO^-), tetralkylammonium (—$N(CH_3)_4^+$), phosphonate (—PO_3H^-), and many others. If the polymer is sufficiently hydrophobic that it will not dissolve in water, it may be formed into a film without further processing, but more hydrophilic polymers may require cross-linking to avoid excessive water solubility.

As a result of their ionic charge, ion-exchange membranes always contain counterions (gegen-ions) that have the opposing charge, and they maintain electrical neutrality. On contact with water, the membranes swell as the ions are hydrated, and they frequently have sufficient water in their structures to form a second continuous phase. This water is beneficial, since it allows ionic mobility and, by plasticizing the membrane polymers, some polymer mobility also. The water-swollen membrane has more space between adjacent polymer chains (greater free volume), and this allows the polymer chains to adjust their positions to allow ions to pass.

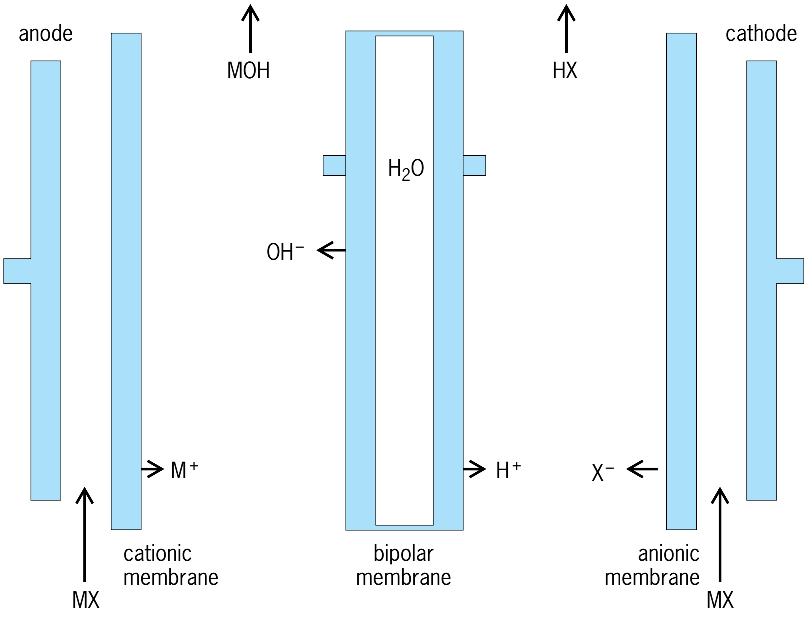


Fig. 4. Bipolar membrane salt-splitting process; water within the membrane splits into hydrogen (H^+) and hydroxide (OH^-) ions, which combine with the anions and cations to form acid and alkaline products.

The Donnan effect is important to the operation of ion-exchange membranes. The Donnan equilibrium may be written for the simple case of univalent ions [Eq. (1)], where the o subscripts indicate

$$[M^+]_o[X^-]_o = [M^+]_m[X^-]_m = [MX] \qquad (1)$$

the solution outside the membrane; and the m subscripts indicate the interior of the membrane; the bracketed terms indicate the equilibrium concentrations. Since the membrane contains a large number of fixed charges, the concentration of counter-ions is also high, and the concentration of co-ions is therefore low. Thus an ion-exchange membrane containing negative fixed charges will readily permit transport of positive counterions, but it will act as a barrier to negative co-ions. The driving force for this transport may be a concentration gradient or an applied electric field. *See* DONNAN EQUILIBRIUM.

Membrane separators. Ion-exchange membranes are frequently used as membrane separators in electrochemical processes. Important among these are fuel cells, batteries, and chloralkali cells. In these applications, the membrane must be a selective conductor for ions, allowing the circuit to close. It thus separates the two half cells, while allowing the passage of ions necessary for the overall process. *See* FUEL CELL.

An example is the process for manufacturing chlorine. A sodium chloride brine is fed to the anode side of the cell, and an electric current is passed through it. Chloride ions are converted to chlorine gas. Sodium ions pass through the membrane to the cathode side, where water is electrolyzed to hydrogen gas, leaving sodium hydroxide. Since the membrane does not pass chloride ions, the resulting sodium hydroxide is quite pure. *See* ELECTROCHEMICAL PROCESS.

Electrodialysis. Electrodialysis is the process whereby saline or brackish water is rendered potable by separating the ions under the influence of an electric field. It is normally carried out by using a membrane stack (**Fig. 3**), in which cation-exchange membranes and anion-exchange membranes are alternated, and a cathode and anode apply the electric field on opposing sides. The saline solution is fed to all of the intermembrane gaps. Under the influence of the electric field, cations pass through the cation-exchange membranes, and anions through the anion-exchange membranes in opposite directions. The result is that alternate cells are depleted in these ions, while the remaining cells experience a net concentration. The depleted streams form the desalinated product. *See* DIALYSIS; WATER DESALINATION.

Salt splitting with bipolar membranes. Salt splitting is related to electrodialysis in technology, but has the goal of converting a salt (MX) into an acid-base pair [HX + MOH; Eq. (2)] for the univalent case. The key

$$MX + H_2O \rightarrow HX + MOH \qquad (2)$$

added element is a bipolar membrane that has positive charges on one side and negative charges on the other. This membrane is placed between anion- and cation-exchange membranes (**Fig. 4**). In operation, water (H_2O) is split, forming hydrogen (H^+) and hydroxide (OH^-) ions within the bipolar membrane. The hydrogen ions combine with anions passing the anion-exchange membrane to form the acid product, while the products combine with cations passing the cation-exchange membrane to form the alkaline product.

Membrane electrodes. Membrane electrodes are widely used in analytical chemistry, the best-known example being the pH electrode. The key to their success is an ion-exchange membrane that is highly selective for the ion of interest. Membranes may be composed of glass or crystalline materials, or they may be made of polymeric membranes similar to those used in other membrane applications. Liquid membranes are also in common use in this application. The membrane is placed between the solution to be analyzed and a second solution containing the solution of interest. The difference in activity between these two solutions creates an electrical potential that can be measured. High flux is not necessary for this application, but rapid response is essential. *See* ELECTRODE; ION-SELECTIVE MEMBRANES AND ELECTRODES.

Liquid membranes. Liquid membranes differ in configuration from the previous applications. Instead of a solid or polymeric membrane, a liquid phase that is immiscible with the external solution is used. Liquid membranes may be supported, in which case the immiscible liquid wets a porous solid that holds it in place; or they may be unsupported, in which case the membrane consists of emulsion globules dispersed in an external phase. Regardless of configuration, differences in concentration between the two nonmembrane phases create a driving force for transport of the ions across the membrane. In the simplest case, these concentration differences are for the ion of interest. However, it is also possible to drive the transport process by using a concentration difference in another ion linked to the first by equilibrium processes. For example, in hydrometallurgical processes a high concentration of acid on the product side (H_p) of a liquid membrane system has been used to drive the concentration of a metal to this side of the membrane. N. N. Li; S. F. Yates

Bibliography. P. M. Bungay, H. K. Lonsdale, and M. N. de Pinho (eds.), *Synthetic Membranes: Science, Engineering and Applications*, 1983; P. R. Haddad and P. E. Jackson, *Ion Chromatography: Principles and Applications*, 1990; W. S. Hancock, *High Performance Liquid Chromatography in Biotechnology*, 1990; N. N. Li and J. M. Calo, *Separation and Purification Technology*, 1992; P. Meares, *Membrane Separation Processes*, 1976; C. F. Poole and S. K. Poole, *Chromatography Today*, 1991; R. Rautenbach and R. Albrecht, *Membrane Processes*, 1989; O. Samuelson, *Ion Exchange Separations in Analytical Chemistry*, 1963; H. Small, *Ion Chromatography*, 1989; H. F. Walton and R. D. Rocklin, *Ion Exchange in Analytical Chemistry*, 1990.

Ion implantation

A process that utilizes accelerated ions to penetrate a solid surface. The implanted ions can be used to modify the surface composition, structure, or property of the solid material. This surface modification depends on the ion species, energy, and flux. The penetration depth can be controlled by adjusting the ion energy and the type of ions used. The total number of ions incorporated into the solid is determined by the ion flux and the duration of implantation. This technique allows for the precise placement of ions in a solid at low temperatures. It is used for many applications such as modifying the electrical properties of semiconductors and improving the mechanical or chemical properties of alloys, metals, and dielectrics. *See* ALLOY; DIELECTRIC MATERIALS; METAL; SEMICONDUCTOR.

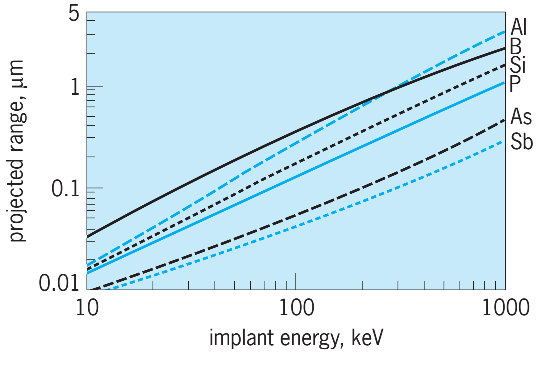
Wide ranges of ion energy and dose are applied. For ion energy ranging from 1 keV to 10 MeV, the ion penetration depth varies from 10 nanometers to 50 micrometers. In general, it is difficult to get deeper penetration since extremely high energy ions are required. As such, ion implantation is a surface modification technique and not suitable for changing the entire bulk property of a solid. Ion dosage also varies depending on the applications. Doses ranging from 10^{10} to 10^{18} ions/cm^2 are typically applied. For high-dose applications, ion sources providing high ion currents are needed to keep the implantation time reasonable for production purposes. *See* ION.

System. An ion implantation system consists of an ion source, an analyzer, an accelerator, a scanner, and a dose integrator at the target. In the ion source, a gaseous mixture is excited to form a plasma containing ions, electrons, and neutral particles. The ions enter the mass analyzer, where they are separated by an adjustable magnetic field for ion selection. The desired ions are chosen based on the charge-to-mass ratio of the ions, and the analyzer is sensitive enough to discriminate against adjacent mass numbers. The ions are then accelerated to the desired ion energy. The ion beam is scanned over the entire solid surface either electrostatically or mechanically. The total ion dose is determined by the charge integrator near the target. When ions are implanted into insulators, an electron source near the solid can be used to supply the surface with electrons to prevent charge buildup. *See* ELECTRON; PARTICLE ACCELERATOR.

Alternative ion sources include focused ion beam systems and immersed plasma ion systems. In focused ion beam sources, ion lenses are used to form a convergent beam of ions with a small spot. Such sources have the flexibility of depositing ions in the desired areas without the use of a masking layer. Since the spot size can be focused down to 10 nm, selective ion implantation is possible with very high resolution. The drawbacks for focused ion beam sources are long processing times (since it is a “direct-write” serial process) and the limited types of available ions. In an immersed plasma ion source, the solid to be implanted is placed inside the plasma where ions are generated. The electric field between the source and the target accelerates the ions to the desired energy. This technique has the advantages of achieving higher ion density, covering a larger area, and allowing solids with different shapes to be implanted. *See* ELECTRIC FIELD; ION SOURCES; PLASMA (PHYSICS).

Projected ion range. As ions enter into a solid, they undergo a number of collisions with the atoms in the solid until they finally come to rest at some depth below the surface, which is the projected ion range. The projected range is proportional to the ion energy and inversely proportional to the ion mass. The projected ranges (see **illus.**) for different ion species implanted into silicon increase proportionally to the energy (10 keV–1 MeV). For boron implanted in silicon, which is often used as a *p*-type dopant for transistors, a high ion energy of 400 keV is required to obtain a 1-μm-deep projected range. Much higher ion energy is needed to achieve projected ranges beyond 1 μm, and it is difficult and expensive to accomplish. *See* BORON; SILICON.

The interactions between the implanted ions and the solid include elastic collisions between pairs of nuclei and inelastic collisions with electrons in the solid. High-energy ions initially lose their energy mostly by electron collisions and travel in a nearly straight path. As energy is reduced, nuclear collisions start to dominate at low energy and a zigzag path begins. Ions stop in the solid when ion energy is decreased to zero after a large number of collisions and changes of ion path. The projected range is the average depth of the implanted ions. The ion distribution can be approximated as a gaussian distribution with a standard deviation. The concentration of the implanted ions is lower near the surface, peaks at the projected range, and decreases beyond the projected range. The ion dispersion occurs in the vertical as well as the horizontal directions. Typically, ion dispersion decreases as the ion energy or the mass of the implanted ion increases. *See* ION-SOLID INTERACTIONS.



Projected range as a function of implant energy for ions (aluminum, boron, silicon, phosphorus, arsenic, and antimony) in silicon. (*After J. F. Gibbons, ed., Projected Range Statistics, 2d ed., 1975*)

In crystalline materials, such as single-crystal semiconductors, the atoms are arranged in regular patterns. Ions moving in certain directions can encounter rows or planes of atoms that form long-range open spaces through which the ions can travel without significant collision. Ions can travel down these channels to a much deeper distance, which is called ion channeling. For ions inside a channel, the major energy loss is due to electronic scattering since they have very few collisions with the nuclei. Since the stopping power for electronic scattering is low, the ion channeling path can be much deeper, showing up as a long tail to the implanted distribution. To avoid ion channeling and keep the ion implantation shallow, the angle between the incoming ions and the solid can be tilted, or the ions can be implanted through an amorphous layer. Ion implantation into a silicon wafer is typically carried out by tilting the solid 7° to minimize ion channeling. *See* CHANNELING IN SOLIDS.

Ion damage and annealing. As ions are implanted into a solid, they undergo a series of nuclear collisions. Every time an ion is scattered, part of the energy is transferred to the atoms in the solid, possibly displacing atoms in its path. In addition, the implanted ions generally do not rest in the proper sites, and implant-induced damage often occurs (due to the impact by high-energy ions). This damage effect is especially critical for crystalline solids. Thermal annealing is used both to repair the crystalline defects produced by the implanted ions and to allow the ions to move to their proper sites (so that the desired surface modifications can be realized). Annealing at temperature above 600°C (1112°F) for a few minutes is usually needed to remove the ion damage. In some cases, it may be helpful to anneal the damaged solid to cause surface melting, providing regrowth by liquid phase. During thermal annealing, diffusion (or distribution) of the implanted ions can occur. To avoid ion diffusion, rapid thermal annealing is used to remove damage at higher temperatures, but only by using a very short annealing time of a few seconds.

Applications. Ion implantation is used extensively in the semiconductor industry. The fabrication of integrated circuits in silicon often requires many steps of ion implantation with different ion species and energies. The implanted ions serve as dopants in semiconductors, changing their conductivity by more than a factor of 10^8. To introduce more electrons in silicon, *n*-type dopants such as arsenic, phosphorus, and antimony are often used, whereas boron is usually used as a *p*-type dopant to provide more holes in silicon. By implanting high doses of oxygen ions in silicon, an insulating oxide layer can be formed underneath the surface, providing a silicon-on-insulator (SOI) structure. These SOI wafers can be used to produce integrated circuits with high speed and power. *See* ANTIMONY; INTEGRATED CIRCUITS; PHOSPHORUS.

Ion implantation is also used to change the surface properties of metals and alloys. It has been applied successfully to improve wear resistance, fatigue life, corrosion protection, and chemical resistance of different materials. Even though the ion projected range is less than 1 μm, surface treatment by ion implantation can extend the lives of metal or ceramic tools by 80 times or more. Ion implantation can form new compounds such as nitrides on the surface, and the implanted ions can be found at much greater depths than the projected range due to diffusion or mechanical mixing. *See* CERAMICS.

Stella W. Pang

Bibliography. J. F. Gibbons (ed.), *Projected Range Statistics*, 2d ed., 1975; F. Komarov, *Ion Beam Modification of Metals*, 1992; J. W. Mayer (ed.), *Ion Implantation in Semiconductors, Silicon, and Germanium*, 1970; E. Rimini (ed.), *Ion Implantation: Basics to Device Fabrication*, 1994; J. F. Ziegler (ed.), *Handbook of Ion Implantation Technology*, 1992; J. F. Ziegler (ed.), *Ion Implantation: Science and Technology*, 1988.

Ion propulsion

Vehicular propulsion caused by the high-speed discharge of a beam of electrically charged minute particles. These particles, usually positive ions, are generated and accelerated in an electrostatic field produced within an ion thruster attached to a spacecraft. Because positive ions cannot be ejected from the thruster without leaving a substantial negative charge on the thruster and spacecraft, electrons must be ejected at the same rate. Ion propulsion systems are attractive because they expel the ions at very high speeds and, therefore, require much less propellant than other thrusters, such as chemical rockets.

The three principal components of an ion propulsion system (**Fig. 1**) are the power-generation and -conditioning subsystem, the propellant storage and feed subsystem, and one or more ion thrusters.

Power generation. This portion of the system can be broken down into power source and power-conversion and -conditioning equipment.

Power source. The power source can be a nuclear reactor or a radiant-energy collector. In the former, thermal power is released by fission or fusion reactions. Mass-minimization considerations generally define spacecraft subsystem configurations. The dominant masses in the case of the nuclear power source are generally those for the main radiator required to reject waste heat from the thermodynamic cycle, and shielding required to protect equipment and personnel against ionizing radiation from the reactor.

Solar radiation as the power source does not require shielding. It can be used to provide electric power directly through photovoltaic (solar) cells or indirectly through a solar collector–heat exchanger system similar to that for a nuclear system (Fig. 1). A single thermodynamic-cycle working fluid suffices, but power levels are generally considered limited to 100 kW by collector-size limitations. Radiant energy beamed to a spacecraft at nonoptical frequencies (for example, as microwaves) and converted to electrical power in appropriate photovoltaic cells may also be attractive in some applications.

Solar cells are a simple and useful power source for propelling small probes into geolunar space, the inner solar system, and in some instances as far as Jupiter. Their principal disadvantage is low voltage. Power-conditioning equipment, to achieve the high voltage required for ion propulsion, adds to the system mass, although considerable advances continue to be made in reducing the complexity and specific mass (mass per unit power) of solar cells and

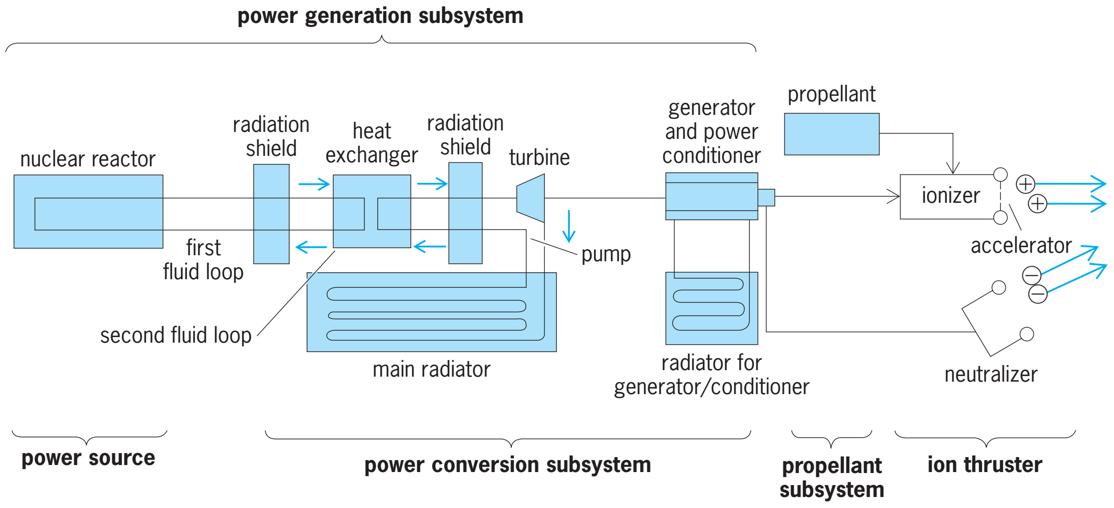


Fig. 1. Nuclear-powered electrostatic (ion) propulsion system.

power-conditioning equipment. Solar cells are damaged by exposure to particle radiation environments, which limits their useful lifetimes to values that may be tolerable in the Earth's radiation belt but are probably too low in the harsher environments at Jupiter or near the Sun.

As solar-powered generators approach the Sun, the energy flux density incident upon their collectors increases, their specific mass decreases, and they become increasingly attractive. Unfortunately, the range of distances from the Sun over which they are truly attractive is not large, because of particle radiation close to the Sun. Additional disadvantages of solar energy systems include the requirement to maintain a collector–concentrator alignment relative to the Sun, the loss of power in the shadow of a celestial body, and the difficulty of collector maintenance in an environment containing micrometeoroids. *See* SOLAR CELL; SOLAR ENERGY.

Power conversion. If the power-generation system involves a nuclear reactor or a solar-thermal subsystem, thermal-to-electric conversion subsystems are required. Those most highly developed involve thermodynamic conversion cycles based on turbine generators. Although most traditional systems have operated on the Brayton gas cycle or the Rankine vapor cycle, more recent efforts include the Stirling gas-cycle system.

In the Rankine cycle, a liquid is heated and evaporated either in a heat exchanger (two-loop system; Fig. 1) or directly in a nuclear reactor or solar-heated boiler (single-loop system). The vapor is then expanded through the turbine, condensed in the main radiator, and pumped back into the heat exchanger, reactor, or boiler. Electrical power is drawn from the generator which is driven by the turbine. Where there is concern about radioactivation of the working fluid (low-melting-temperature metals typically), the two-loop system is used and the heat exchanger serves to limit the spread of radioactivity. Heat transfer is complicated by the weightless conditions of space. Other thermodynamic cycles such as the Brayton or Stirling cycles, which do not require vaporization and condensation, mitigate this problem, but they typically have lower efficiencies. *See* BRAYTON CYCLE; GAS TURBINE; GENERATOR; RANKINE CYCLE; STIRLING ENGINE.

The main radiator, which rejects the waste heat from the conversion subsystem to space, is the largest and heaviest of the power-generation-subsystem components. Its size and mass are directly proportional to the thermal power rejection rate. High efficiency is critical to low radiator mass, but operating temperature, physical configuration, and micrometeor protection are also important.

Alternative methods of power conversion employ thermoelectric (semiconducting thermocouples) and thermionic devices to effect direct conversion of heat to electricity. These devices have a history of lower efficiencies than fluid cycle systems operating between the same temperatures, but they are typically more compact and, because they typically have no moving parts, they are potentially more reliable. *See* THERMIONIC POWER GENERATOR; THERMOELECTRICITY.

Propellants. Ion-thruster propellants are selected according to performance, technical suitability, availability, cost, and storage characteristics. For geospace transportation, the effect of ions that may spiral along the Earth's magnetic field lines to the magnetic poles must also be considered.

High thruster efficiency requires a high-molecular-mass propellant which facilitates operation at high utilization efficiency where the loss of unaccelerated propellant is small. A high molecular mass also facilitates operation at a high acceleration voltage, which assures a high thruster electrical efficiency. Low first-ionization potential and high first-excitation potential for the propellant are important. Such potentials assure minimal radiation losses and low ion-production power.

High thrust-to-power ratio is also promoted by a high propellant molecular mass. Operation at this condition is desirable because spacecraft are usually power-limited.

Thruster lifetime can be extended by a high second ionization potential and a low charge-exchange cross section to limit the rate at which doubly charged and charge-exchange ions erode thruster surfaces.

Compatibility of thruster with spacecraft materials and functions is essential. Propellants that are, for example, likely to interfere with the operation of the thruster or spacecraft systems by forming coatings are undesirable.

Operational controllability is important. Monatomic propellants are preferred because they ionize in a predictable way and the thrust created by their expulsion is, therefore, more readily controlled.

Propellants that have been investigated include argon, xenon, cesium, mercury, and fullerenes such as C_{60} (see **table**). Although mercury received most of the early attention, xenon is now being used on all space missions because of toxicity concerns with mercury. Cesium is not used because it tends to migrate and coat surfaces on components such as insulators. C_{60}, which is attractive because of its high molecular mass, is probably unacceptable because it tends to fragment and to form negative ions, which degrade thruster performance.

Thrust device. Ion or electrostatic thrust devices contain three functional elements: an ionizer that generates the ions; an accelerator providing an electric field for accelerating the ions and forming them into a well-focused beam; and a neutralizer or electron emitter that neutralizes the electrical charge of the exhaust beam of ions after they have been ejected. Electrostatic thrusters are classified according to the type of ions they eject and by the method utilized to produce them.

Ionizer. The positive ions needed for acceleration are produced in a strong electric field, by contact with a surface having a work function greater than the ionization potential of the propellant, or by electron-bombardment ionization. The last method

Ion propellant characteristics

Propellant	Atomic weight	Availability	Storability	Spacecraft thruster material compatibility	Ionization potentials, eV		Thruster technology status
					1st	2d	
Argon	39.95	Good	Cryogenic liquid	Excellent	15.76	27.5	Ground tests
Xenon	131.3	Scarce	High-pressure gas	Excellent	12.13	21.2	Space flight
Cesium	132.9	Limited	Solid	Poor	3.90	23.4	Space flight
Mercury	200.6	Good	Liquid	Good	10.44	18.8	Space flight
Fullerene	720	Good	Solid	Marginal	7.6	?	Preliminary tests

has received the most attention and appears to be the most promising.

In an electron-bombardment ionizer, high-energy electrons collide with relatively massive neutral atoms or heavy molecules like fullerenes being supplied in the discharge chamber and induce ionization (**Fig. 2**). In the direct-current ionizer, the electrons are accelerated from a cathode to kinetic energies of a few tens of electronvolts via a potential difference sustained by an anode. These electrons are usually obtained from a hollow cathode which has the long lifetime, high reliability, and low power consumption needed for space operation. Various magnetic-field configurations, which are usually induced by permanent magnets, are employed to confine the electrons away from the anode until after they have expended most of their kinetic energy in collisions with neutral atoms or molecules. The magnetic field also serves to confine and direct the ions produced toward the accelerating electrodes. This class of thruster has been operated in space with cesium, mercury, and xenon propellants.

In another type of electron-bombardment ionizer, electrons that result from the ionization process itself are accelerated in high-frequency electric fields induced within the discharge chamber by external antennae. These electrons acquire sufficient kinetic energy to ionize propellant atoms; they then migrate toward an electrode where they are collected. In contrast to dc ionizers, these use high-frequency rather than direct-current power to accelerate electrons and require no cathode. *See* ION SOURCES.

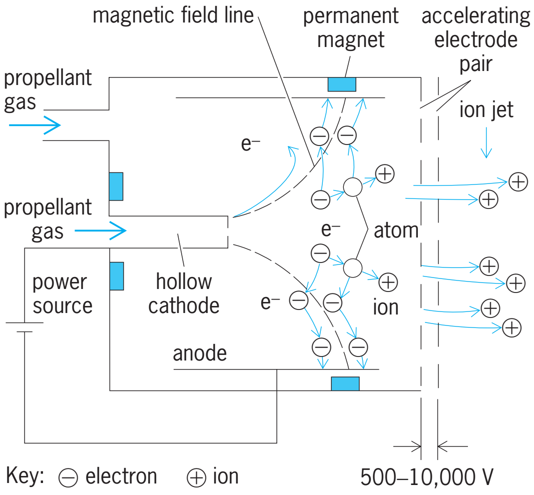


Fig. 2. Direct-current electron-bombardment ion thruster.

Accelerator. Some of the ions produced are directed toward the ion-accelerating subsystem which typically consists of two plates containing large numbers of aligned hole pairs (Fig. 2). The upstream plate and the body of the ionizer are maintained at a potential V (in volts) positive with respect to the space downstream from the thruster, whereas the downstream plate is biased negative at a smaller value. The positive ions are therefore first accelerated between the plates and then decelerated after they pass through the downstream plate until they have been accelerated through the net (or beam) voltage V. In order to ensure that ions do not strike the downstream plate directly, hole size and alignment and plate spacing must be set and maintained properly. For a high extracted ion current density, the plates should be as close together as possible.

Charge neutralization. Electrons are released near the ion beam downstream from the accelerator at the same rate as the ions. These electrons neutralize the space charge of the positive ions and prevent gross spacecraft charging. They are prevented from migrating upstream into the positively biased ionizer by the negatively biased downstream accelerator plate. A hollow cathode, similar to the one used as an electron source in the dc ionizer but optimized for low propellant flow rate, is used as a neutralizer because of its high efficiency and long life.

Stationary plasma thruster. This thruster combines the functions of ionization, acceleration, and neutralization in a simple open chamber without a planar electrode pair. It utilizes a magnetic field to confine electrons while ions are accelerated electrostatically, and current designs have a limited specific impulse capability (1000–2000 s). *See* SPECIFIC IMPULSE.

Applications. Ion propulsion is characterized by high specific impulse and low thrust. Thrusters have generally been designed for operation in the 2500–10,000-s specific impulse range, but the upper limit could be extended. Because high specific impulse means low propellant consumption, ion propulsion is attractive for a wide variety of applications.

Satellites. One functional category includes the use of ion thrusters on satellites for orbit control (against weak perturbation forces) and for station keeping (position maintaining of satellite in a given orbit). Substantial commercial use of ion thrusters in this application began at the end of the twentieth century.

An ion propulsion system can also be used advantageously for changing the satellite's position in a given orbit, especially shifting a satellite to different longitudes over the Earth in an equatorial geostationary orbit. In current systems, the satellite's reaction control system is fired for 2–3 min, causing the satellite to drift slowly to a different place, where it is stopped by a second reaction control system maneuver. The slow drift can be time-consuming, several times the few days required by an ion thruster for a 180° shift. *See* SATELLITE (SPACECRAFT); SPACECRAFT PROPULSION.

Orbit raising and related missions. Ion propulsion in geospace can be advantageous, but it is complicated by three factors: strong gravity forces, which typically result in near-Earth to geosynchronous transfer times of the order of 100 days depending on available power-vehicle mass ratio; the Van Allen radiation belt; and eclipsing. The last two factors are problematical when the Sun's power is used, especially when silicon photovoltaic cells, which suffer radiation damage in the Van Allen belt, are involved. The exclusive use of nuclear power or the use of a reusable high-thrust interorbital stage to boost the solar-electric stage from near-Earth to an orbit above the most damaging region of the Van Allen belt (18,000 km or 11,000 mi) can eliminate the last two concerns. Orbit plane changes accomplished along with orbit raising are also attractive, although they increase transfer times. Geospace missions might involve delivery of cargo, single- and multiple-satellite servicing, and space debris retrieval. This last mission is fast becoming important as the probability of collisions with debris, especially at the geosynchronous altitude, increases. *See* VAN ALLEN RADIATION.

Geolunar transfer. Earth-Moon transfer can be accomplished with ion propulsion starting at any place in the near-Earth to geosynchronous altitude range, but the complicating factors cited above make the higher-altitude starting point more attractive.

Interplanetary transfer. A major functional application of ion propulsion is interplanetary transfer. Here, thrust has to overcome only very weak solar gravitational forces. Because of this, and the long powered flight times of which ion propulsion is capable, transfer times to Venus or Mars need not be longer than transfer times in comparable flights with high thrust drives capable only of short powered flight. At the very large distances to objects in the outer solar system, ion propulsion would yield shorter transfer times than chemical and most high-thrust nuclear concepts. The National Aeronautics and Space Administration (NASA) *Deep Space 1* mission, started at the end of the twentieth century, uses a 30-cm-diameter (12-in.) xenon ion thruster to propel a spacecraft to encounters with the asteroid Braille (in 1999) and the comet Borrelly (2001). *See* ELECTROTHERMAL PROPULSION; INTERPLANETARY PROPULSION; PLASMA PROPULSION; SPACE PROBE.

Paul J. Wilbur

Bibliography. G. R. Brewer, *Ion Propulsion*, 1970; M. S. El-Genk et al., System design optimization for multimegawatt space nuclear power applications, *J. Propuls., Power*, 6:194–202, 1990; H. R. Kaufman, Technology of electron-bombardment ion thrusters, *Adv. Electr. Electr. Phys.*, 36:273–275, 1974; E. J. Lerner, Plasma thrusters from Russia, *Aerosp. Amer.*, 30(9):51, September 1992; H. Mark and H. J. Smith, Fast track to Mars, *Aerosp. Amer.*, 29:36–41, 1991; A. K. Sinha, B. Agrawal, and W. W. Wu, Trends in satellite communications technology, techniques and applications, *Int. J. Satel. Commun.*, 8:283–294, 1990; P. J. Wilbur, J. R. Beattie, and J. Hyman, Approach to the parametric design of ion thrusters, *J. Propuls., Power*, 5:575–583, 1990; P. J. Wilbur, R. G. Jahn, and F. C. Curran, Space electric propulsion plasmas, *IEEE Trans. Plasma Sci.*, 19:1167–1179, 1991; P. J. Wilbur, V. K. Rawlin, and J. R. Beattie, Ion thruster development trends and status in the United States, *J. Propuls., Power*, 14:708–715, 1998.

Ion-selective membranes and electrodes

Membrane-based devices, involving permselective, ion-conducting materials, used for the measurement of activities of species in liquids or partial pressures in the gas phase. Permselective means that ions of one sign may enter and pass through a membrane.

Properties. Ion-selective electrodes are generally used in the potentiometric mode, and they superficially resemble the classical redox electrodes of types 0 (inert), 1 (silver/silver ion; Ag/Ag^+), 2 (silver/silver chloride, chloride ion; Ag/AgCl/Cl-, and 3 (lead/lead oxalate/calcium oxalate/calcium ion; $Pb/PbC_2O_4CaC_2O_4/Ca^{2+}$). The last, while ion-selective, depends on a redox couple (electron exchange) rather than ion exchange as the principal origin of interfacial potential difference. Ion-selective electrodes have a typical form that can be expressed in shorthand form, the "ionic-contact" membrane configuration: [lead wire (Cu; Ag); inner reference electrode (AgCl); inner filling solution (electrolyte: M^+Cl^-); membrane permselective to M^+] and the "all-solid-state" membrane configuration [Cu; Ag membrane permselective to M^+].

In the former, both membrane interfaces are ion-exchange-active, and the potential response depends on M^+ activities in both the test solution and the inner filling solution. In the latter, the membrane must possess sufficient electronic conductivity to provide a reversible, stable electron-exchange potential difference at the inner interface, with ion exchange only at the test solution side.

Potentiometric responses of ion-selective electrodes take the form in Eq. (1) when an ion-selective

$$V(\text{measured}) = V^0 + \frac{RT}{F}\ln\left[a_i^{1/z_i} + \sum_j (k_{ij}a_j)^{1/z_j}\right] \qquad (1)$$

$R =$ the universal gas constant
$T =$ the absolute temperature
$F =$ the Faraday constant (96,487 coulombs/equivalent)
$V_0 =$ formal reference potential

electrode is used with an external reference electrode, typically a saturated-calomel reference electrode with a salt bridge, to form a complete electrochemical cell. Activities of the principal ion, a_i, and interfering ions, a_j, are in the external "test" solution and correspond to ions M^{zi} and M^{zj}, where z_i and z_j are charges with sign. The ion M^{zj} is written first because it is the principal ion favored in the membrane, for example, high ion-exchange constant and high mobility. Values known as selectivity coefficients (k_{ij}) are experimentally determined, but they can be related to extrathermodynamic quantities such as single-ion partition coefficients and single-ion activity coefficients and mobilities. When only one ion is present in a test solution, or the membrane is ideally permselective for only one ion, this equation simplifies as Eq. (2). V^0 can be

$$V(\text{measured}) = V^0 + \frac{RT}{z_i F} (\ln a_i) \qquad (2)$$

written explicitly in terms of activities of species at the inner interface or in terms of solid-state activities for the all-solid-state configuration. Equation (1), variously known as the Horovitz, Nicolsky, or Eisenman equation, resembles the Nernst equation (2), but the former originates from different factors. Equation (1) cannot be derived from first principles for ions of general charge. However, when $z_i = z_j$, the equation can be derived by various means including, mainly, thermodynamic (Scatchard) equations and transport models (Nernst-Planck equations).

The response slope $2.303RT/z_iF$ is $59.14/z_i$ mV per decade of activity at 77°F (25°C). Measurements reproducible to ± 0.6 mV are typically achieved, and activities can be reproducible to ± 2% for monovalent ions. A normal slope is considered nernstian, and can persist over a wide activity range, especially for solid electrodes, for example, 24 decades in the Ag^+,S^{2-} system, and 12 decades of H^+ using Li^+-based glass membrane electrodes. Less than nernstian slopes, and in the limit zero slope, can occur at low activities of sensed species, and can occasionally occur at very high activities. Ultimate low-level response (detection limit) is determined by the solubility of the membrane ion-exchanging material, although impurities may cause premature failure. Because of the logarithmic dependence of potential response on activities, activity measurements using ion-selective electrodes are particularly suited to samples with wide activity variations. Standardization against pure samples or samples of the type to be determined is required. Precise measurements of concentrations over narrow ranges are not favorable, but are possible by elaborate standardization schemes: bracketing, standard additions, and related methods involving sample pretreatment. *See* DONNAN EQUILIBRIUM.

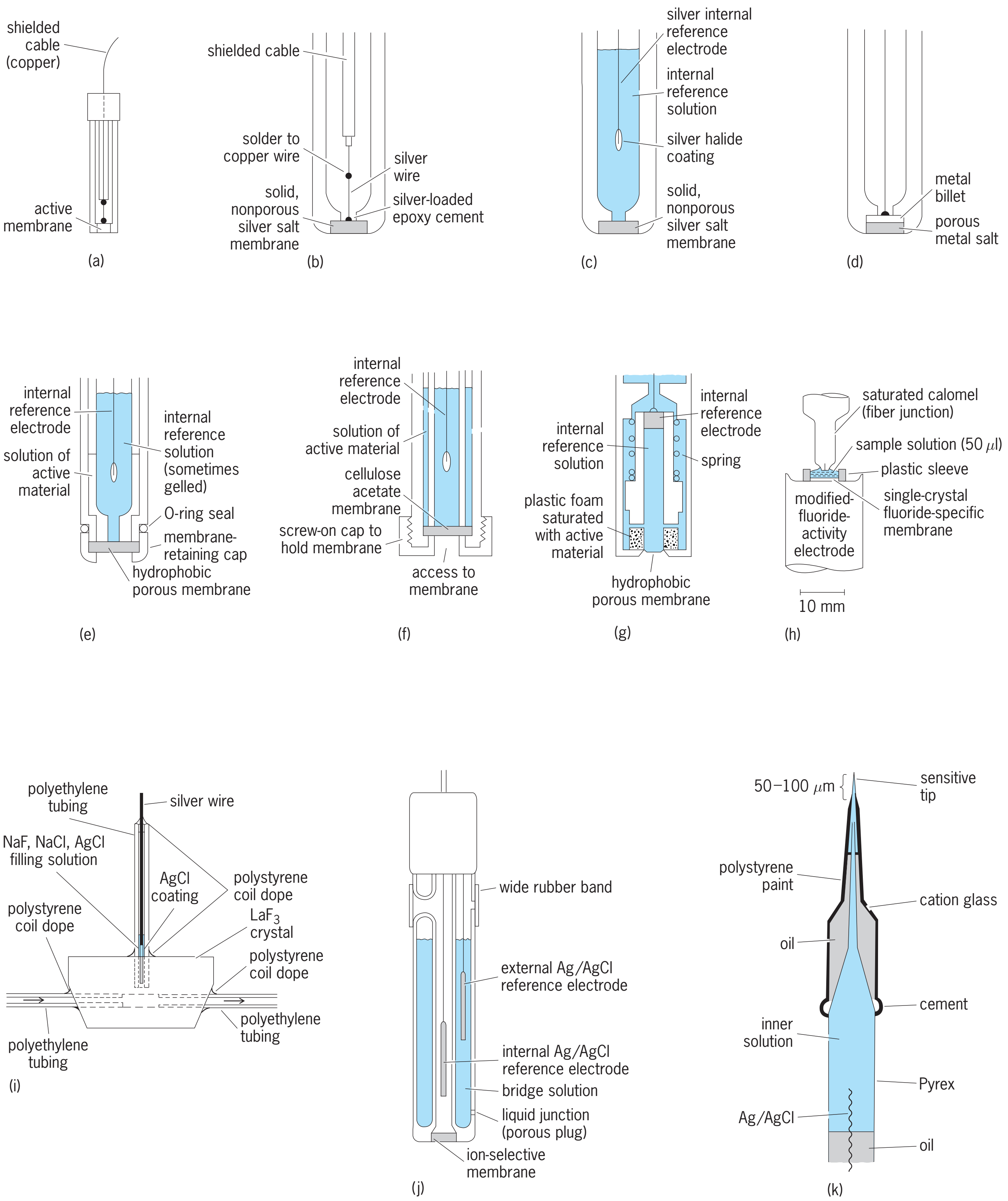
Ion-selective electrodes are most often cylindrical, 6 in. (15 cm) long and 0.25 or 0.5 in. (6 or 13 mm) in diameter, with the lead wire existing at the top and the membrane sensor at the lower end. However, inverted electrodes for accommodating single-drop samples, and solid electrodes with a drilled hole or a special cap for channeling flowing samples past a supported liquid membrane, are possible configurations. The conventional format is intended for dip measurements with samples large enough to provide space for an external reference electrode. Single combination electrodes are useful for smaller samples, because the external electrode is built-in nearly concentrically about the ion-selective electrode. Drilled and channeled-cap electrodes are intended for use with flowing samples. Microelectrodes with membrane-tip diameters of a few tenths of a micrometer have been constructed for single-cell and other measurements in the living body (see **illus.** for construction details).

Electrodes can also be flat-form stacks of active layers on an inert support, that is, Ag/AgCl/KCl (aqueous gel)/membrane. Membranes can also be placed on an inert metal gate or insulator of a field-effect device. These flat devices are susceptible to miniaturization by methods of silicon chip technology. So-called smart sensors have the signal-processing electronics on the chip as well.

Ion-selective electrodes are intended to be used to monitor and measure activities of flowing, or stirred, solutions because electrodes detect and respond to activities only at their surfaces. The time responses of solid and liquid membrane electrodes to ideal step activity changes of the principal ion (already present in the membrane) can be very rapid: 200 milliseconds for glass to about 30 microseconds for silver bromide (AgBr). Generally this fast response cannot be observed or used, because sample mixing or diffusion of a fresh sample to an electrode surface is the limiting process. Also, almost any time two ions are simultaneously determining response, interior diffusion potential generation is involved in reaching a new steady-state potential. Similarly, formation of hydrated surface layers or layers of adsorbed matter or inhomogeneities introduces diffusion barriers. Response times from 2 to 20 s can be expected. About 20–30 samples per hour may be analyzed manually by the dip method, and about 60 per hour when samples are injected into a flowing stream of electrolyte. *See* ELECTRODE POTENTIAL; ELECTROLYTIC CONDUCTANCE.

Classification and responsive ions. Ion-selective electrodes are classified mainly according to the physical state of the ion-responsive membrane material, and not with respect to the ions sensed. It has also proved superfluous to distinguish between homogeneous membranes and those that are made from a homogeneous phase supported physically in voids of an inert polymer, or from two homogeneous phases intimately mixed, so-called heterogeneous membranes.

Glass membrane electrodes. These are used for hydrogen-ion activity measurements. Glass electrodes are based on alkali ion silicate compositions. Superior pH-sensing glasses (pH 1 to 13 or 14) result from lithium silicates with addition of di-, tri-, and tetravalent heavy-metal oxides. The latter are not chain formers. Membranes responsive to



Diagrams of ion-selective electrodes. (*a*) Typical electrode configuration—in this example, an all-solid-state ion-selective electrode. (*b*) Enlarged view of the construction for metal-contacted-membrane ion-selective electrodes. (*c*) Enlarged view of the construction for internal electrolyte-contacted-membrane ion-selective electrodes. (*d*) Enlarged view of the construction for an electrode. (*e–g*) Enlarged views of constructions for liquid ion-exchanger-membrane ion-selective electrodes. (*h*) An inverted electrode microcell using a fluoride-sensing material; the reference electrode (external) is saturated calomel [SCE] (*after R. A. Durst and J. K. Taylor*, *Anal. Chem.*, *39*:*1483*, *1967*). (*i*) Construction of a flow-through crystal electrode (*after H. I. Thompson and G. A. Rechnitz*, *Anal. Chem.*, *44*:*300*, *1972*). (*j*) A combination electrode illustrating the usual active membrane surrounded by an attached Ag/AgCl external reference electrode. (*k*) An example of a cation-sensing microelectrode used in biological research (*after R. N. Khuri*, *W. J. Flanagan*, *and D. E. Oken*, *J. Appl. Physiol.*, *21*:*1568*, *1966*).

the sodium (Na^+), potassium (K^+), and ammonium ($NH_4{}^+$) cations, and some other cations use additional aluminum oxide (Al_2O_3), boric oxide (B_2O_3), or both. The pH glasses are highly selective for hydrogen ion (H^+) over other monovalent ions. The Na^+-sensing glasses are not intrinsically very selective for Na^+ over H^+, but useful pNa measurements can be made, even in excess K^+ at pH 7 or above. No glasses with high selectivity of K^+ over Na^+ have been found. Chalcogenide glasses containing low contents of copper(III) [Cu^{2+}] or iron(III) [Fe^{3+}], while called glasses, are thought to be semiconductor electrodes with a high component of electron exchange, rather than ion exchange, for establishment of interfacial potential responses to Cu^{2+} and Fe^{3+}.

Electrodes based in water-insoluble inorganic salts. These electrodes include sensors for the ions fluoride (F^-), chloride (Cl^-), bromide (Br^-), iodide (I^-), cyanide (CN^-), thiocyanate (SCN^-), sulfide (S^{2-}), Ag^+, Cu^{2+}, cadmium (Cd^{2+}), and Pb^{2+}. The compounds used are silver salts, mercury salts, sulfides of Cu, Pb, and Cd, and rare-earth salts. All of these are so-called white metals whose aqueous cations (except lanthanum; La^{3+}) are labile. The salts are Frenkel-defect solids which possess the necessary ionic conductivity. $Ag_i{}^+$ (interstitials) or $Ag_i{}^-$ (vacancies) are the mobile species in the silver salts, while F^- interstitials are mobile in LaF_3. These materials are ion exchangers, and show no diffusion potential. Single crystals, doped and undoped, may be used as membranes. Pressed pellets using inert binders such as polyethylene or an insoluble salt such as Ag_2S (for the silver halide electrodes) are popular.

In addition, powdered salts may be suspended in silicone rubber or polyvinyl chloride (about 50:50% by weight) to form heterogeneous flexible membranes. CuS-Ag_2S, CdS-Ag_2S, and PbS-Ag_2S pressed pellets formed at about 480°F (250°C) are indirectly responsive to the divalent metal ion activities through control of Ag^+ activities at the electrode surface and in leached layers or surface pores by means of the common ion effect.

Electrodes using liquid-ion exchangers. These are electrodes supported in the voids of inert polymers such as cellulose acetate, or in transparent films of polyvinyl chloride, and provide extensive examples of devices for sensing. Fewer cation-sensing liquid-ion exchanger systems have been found. The principal example (and among the most important) is the Ca^{2+}-responsive electrode based on calcium salts of diesters of oil-soluble phosphonic acids. Oil-soluble metal salts of hydrophobic acids containing chelating nitrogen (N) and sulfur (S) groups can be used in electrodes for heavy-metal ion sensing. The condition for development of an ion-selective potential signal is rapid exchange of the sensed metal ion between membrane (dissolved chelate or salt) and the same ion in the test solution. If this exchange is slow, as in magnesium ion (Mg^{2+}) exchange between porphyrins and solutions, the sensor is anion-selective. Anion-sensing electrodes typically use an oil-soluble cation Aliquat (methyltricaprylammonium) or a metal ion-uncharged organic chelating agent (Ni^{2+} or Fe^{2+} phenanthroline or substituted phenanthroline cations) in a support matrix. Sensitivity is virtually assured if the salt is soluble in a mediator solvent, typically a nitro aromatic or esters of difunctional carboxylic acids: adipic, sebacic, or phthalic. Selectivity poses a severe problem since these electrodes, based on hydrophobic materials, tend to respond favorably to many oil-soluble anions. Thus construction of electrodes for the simple inorganic anions F^-, hydroxide (OH^-), bicarbonate ($HCO_3{}^-$), and hydrogen phosphate ($HPO_4{}^{2-}$) is difficult. Yet many electrodes respond to SCN^-, I^-, Br^-, nitrate ($NO_3{}^-$), perchlorate ($ClO_4{}^-$), and fluoroborate ($BF_4{}^-$) in accordance with the Hofmeister lyotropic series. Surfactant anion sensors use salts such as hexadecylpyridinium dodecylsulfate in *o*-dichlorobenzene; surfactant cation sensors use a picrate salt of the species to be measured. Acetylcholine may be measured in the presence of choline, Na^+, and K^+ using the tetra-*p*-chlorophenylborate salt in a phthalate ester in polyvinyl chloride, for example. Cation and anion drug sensors can be made by dissolving a cation pair (with oil-soluble anion such as tetraphenylborate) in a plasticized polyvinyl chloride membrane. Anion drug sensors often use methyltridodecylammonium ion (Aliquat) to form the oil-soluble ion pair.

Neutral carrier-based sensors for monovalent and divalent cations are closely related to ion-exchanger-based electrodes. Both systems involve ion-exchange sites, particularly negative mobile sites arising from mediators and negative fixed sites arising from support materials. Some of the available neutral carriers are hydrophobic complex formers with cations, and they can be either cyclic or open-chain species that form ion-dipole complexes. These compounds permit selective extraction (leading to permselectivity) for ions such as K^+, Na^+, $NH_4{}^+$, and Ca^{2+} that would ordinarily not dissolve as simple inorganic salts in the hydrocarbonlike membrane phase. Valinomycin is the best-known example, and its use in supported solvents such as dioctylsebacate provides an electrode with sensitivity of 10^5 for K^+/Na^+. Neutral carriers that form adducts using weak covalent bonds are particularly useful for constructing selective anion sensors without Hofmeister series limitations. Trifluoroacetyl-*p*-butylbenzene$_2$, selective for carbonate ($CO_3{}^{2-}$); bis(*p*-chlorobenzyl)tin dichloride, selective for $HPO_4{}^{-2}$; and some special porphyrins are examples. *See* ION EXCHANGE.

Electrodes with interposed chemical reactions. These electrodes, with chemical reactions between the sample and the sensor surface, permit an enhanced degree of freedom in design of sensors for species which do not directly respond at an electrode surface. Two primary examples are the categories of gas sensors and of electrodes which use enzyme-catalyzed reactions. Gas sensors for carbon dioxide (CO_2), sulfur dioxide (SO_2), ammonia (NH_3), hydrogen sulfide (H_2S), and hydrogen chloride (HCl), and others can be made from electrodes responsive to H^+, S^{2-}, or Cl^-. By enclosing a pH glass membrane in a thin layer of dilute sodium bicarbonate ($NaHCO_3$), an electrode for partial pressure of CO_2 is formed, since H^+ increases in a known way with increasing dissolved CO_2. Similarly, immobilized enzymes convert a substrate such as urea or an amino acid to ammonia, which can be sensed and monitored by the underlying electrode. However, increased sensitivity is accompanied by an increased response time. Each diffusion and diffusion-reaction barrier slows the transport and increases the time constant of the overall sensor electrode. *See* ELECTRODE; ENZYME.

Applications. Electrodes for species identified above are, for the most part, commercially available. In addition, electrodes have been made and reported that are responsive to many other species. A few of these are cesium (Cs^+), thallium (Tl^+), strontium (Sr^{2+}), Mg^{2+}, zinc (Zn^{2+}), nickel (Ni^{2+}), (UO_2^{2+}), mercury(II) [Hg], hydrogen sulfite (HSO_3^-), sulfate (SO_4^{2-}), periodate (IO_4^-), perrhenate (ReO_4^-), halide anion complexes of heavy metals (for example, $FeCl_4^-$), pyridinium, pyrocatechol violet, vitamins B_1 and B_6, and many cationic drugs, aromatic sulfonates, salicylate, trifluoroacetate, and many other organic anions. Applications may be batch or continuous. Important batch examples are potentiometric titrations with ion-selective electrode end-point detection, determination of stability constants of complexes and speciation identity, solubility and activity coefficient determinations, and monitoring of reaction kinetics, especially for oscillating reactions. Ion-selective electrodes serve as liquid chromatography detectors and as quality-control monitors in drug manufacture. Applications occur in air and water quality (soil, clay, ore, natural-water, water-treatment, seawater, and pesticide analyses); medical and clinical laboratories (serum, urine, sweat, gastric-juices, extracellular-fluid, dental-enamel, and milk analyses); and industrial laboratories (heavy-chemical, metallurgical, glass, beverage, and household-product analyses). *See* ANALYTICAL CHEMISTRY; CHROMATOGRAPHY; QUALITY CONTROL; TITRATION. Richard P. Buck

Bibliography. P. L. Bailey, *Analysis with Ion-Selective Electrodes*,1988; J. Koryta and K. Stulik, *Ion-Selective Electrodes*, 2d ed., 1984; E. Pungor (ed.), *Ion-Selective Electrodes* vol. 5, 1989; R. L. Solsky, *Analytical Chemistry*, *Fundamental Reviews*, vol. 60, pp. 106R–113R, 1988.

Ion-solid interactions

Physical processes resulting from the collision of energetic ions, atoms, or molecules with condensed matter. These include elastic and inelastic backscattering of the projectile, penetration of the solid by the projectile, emission of electrons and photons from the surface, sputtering and desorption of neutral atoms and ions, production of defects in crystals, creation of nuclear tracks in insulating solids, and electrical, chemical, and physical changes to the irradiated matter resulting from the passage or implantation of the projectile (**Fig. 1**). Ion-solid interactions are also known as particle-solid interactions.

When an energetic ion which has been produced in a plasma, a particle accelerator, or an astrophysical process impinges upon the surface of condensed matter, it experiences a series of elastic and inelastic collisions with the atoms which lie in its path. These collisions occur because of the electrical forces between the nucleus and electrons of the projectile and those of the atoms which constitute the solid target. They result in the transformation of the energy of the projectile into internal excitation of the solid. The precise nature of this excitation and the resulting physical processes are determined largely by the bombarding conditions. The principal determining factors are the species of ion, its energy and direction of incidence, the target composition and crystal structure, the target temperature, and the condition of the target surface. Such factors as the quantum-mechanical state of the projectile may also influence specific processes, especially if they are known to occur very near the target surface.

Backscattering. One of the most basic interactions occurs when the projectile collides with a surface atom and bounces back in generally the opposite direction from which it came. This process is known as backscattering. It was first observed in 1911, when Ernest Rutherford bombarded gold foils with alpha particles from radioactive decay, and its observation led Rutherford to conclude that most of the matter in atoms is concentrated in a small nucleus. Now it is used as an analytical technique, Rutherford backscattering analysis, to measure the masses and locations of atoms on and near a surface. By measuring the energy and direction of a backscattered particle whose initial energy and direction are known, the mass of a surface atom that was struck can be inferred. Moreover, when the scattering occurs below the surface, small but predictable shifts in the measured energy can be used to infer the depth of the collision. For projectiles such as alpha particles with energies of a few MeV, the cross section for backscattering is well known. As a result, the fraction of backscattered particles is an absolute measure of the number of targets on the surface. This technique is most commonly performed with alpha particles of about 2 MeV. It is widely used in the study of other ion-solid phenomena and in analyzing the thickness and composition of thin layers, a measurement often of considerable value in research on semiconductor devices. *See* SCATTERING EXPERIMENTS (NUCLEI).

Another backscattering technique, known as ion-scattering spectrometry, uses projectiles with energies of perhaps 2 keV and thus achieves significantly greater surface specificity. However, results are more qualitative since the absolute cross sections for these collisions are not well known. Many of the advantages of Rutherford-backscattering and ion-scattering spectrometries can be combined by the use of ions in the 100–300-keV energy range. This medium-energy backscattering spectrometry has been made possible by advances in time-of-flight particle detection technology.

Penetration phenomena. Although backscattering events are well enough understood to be used as analytical tools, they are relatively rare because they represent nearly head-on collisions between two nuclei. Far more commonly, a collision simply deflects the projectile a few degrees from its original direction and slows it somewhat, transferring some of its kinetic energy to the atom that is struck. Thus, the projectile does not rebound from the surface but penetrates deep within the solid, dissipating its kinetic energy in a series of grazing collisions.

The capacity of a solid to slow a projectile is called the stopping power $S(E)$, and is defined as the amount of energy lost by the projectile per unit length of trajectory in the solid. The stopping power is a function of the projectile's energy and has different values for different projectiles. It is commonly measured in units of eV/cm or an equivalent unit of eV/(g/cm^2) obtained by dividing the former value by the density of the solid. Stopping power is of central importance for many phenomena because it measures the capacity of a projectile to deposit energy within a thin layer of the solid. Since this energy drives secondary processes associated with penetration, the stopping power of the projectile is an important scaling parameter.

In order to understand the graph of the stopping power of an energetic nucleus in matter as a function of its energy (**Fig. 2**), it is necessary to distinguish collisions involving the nuclei and the electrons of the target.

When one charged particle moves past another at high velocity, the average force acting on the stationary particle is perpendicular to the trajectory of the moving particle. This force acts for a brief period of time and gives the stationary particle a small momentum. To simplify calculations, it is assumed that the target particle, whose mass is m, does not move very much during the collision and that the moving particle is not significantly deflected by it. This is called the impulse approximation. The magnitude Δp of the momentum acquired by the struck particle is independent of its mass so that the energy that it receives in the collision is $(\Delta p)^2/2m$. Thus, an electron whose mass is about 1/2000 that of even a single nucleon will receive much more energy from a collision than a nucleus. The portion of the graph (Fig. 2) for which this simple explanation is satisfactory has a slope proportional to $1/E$ and is labeled the Bethe region. Here a projectile penetrates the target as a bare nucleus devoid of electrons because they have been stripped away in collisions.

At relativistic velocities the electric field of the projectile becomes increasingly concentrated in the plane perpendicular to the trajectory where it can interact with the target more efficiently. As a result, the stopping power eventually reaches a minimum, called the ionization minimum, and then begins to rise again.

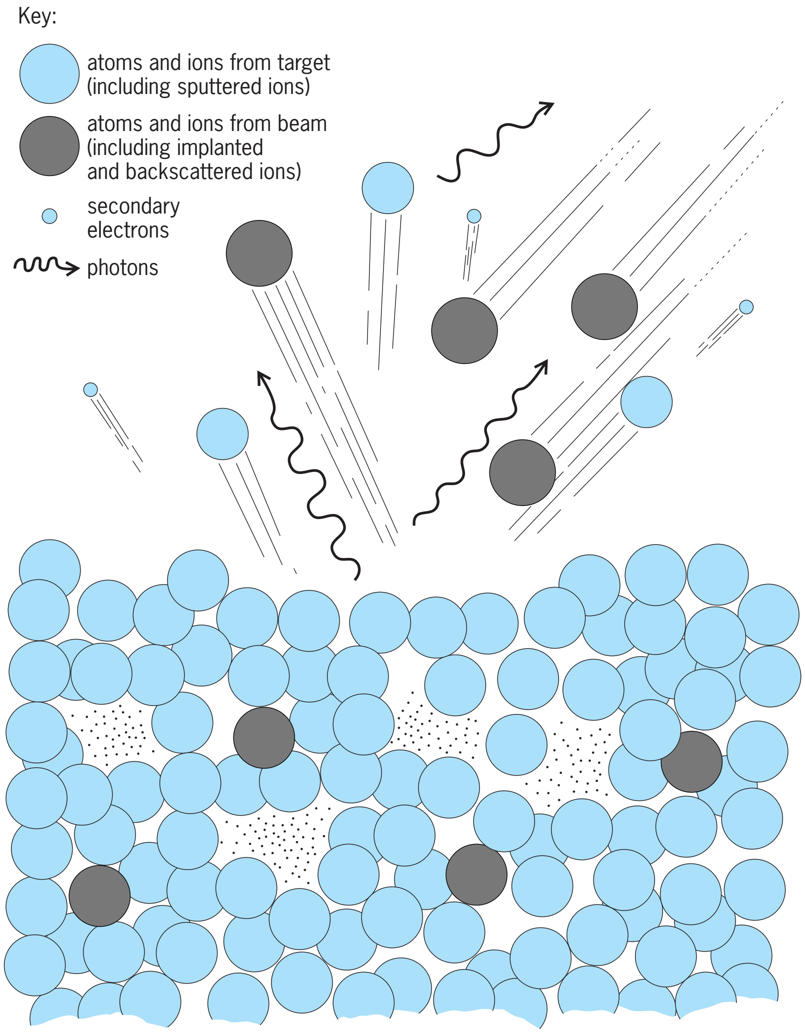


Fig. 1. **Processes that can occur when a material is subjected to particle bombardment.**

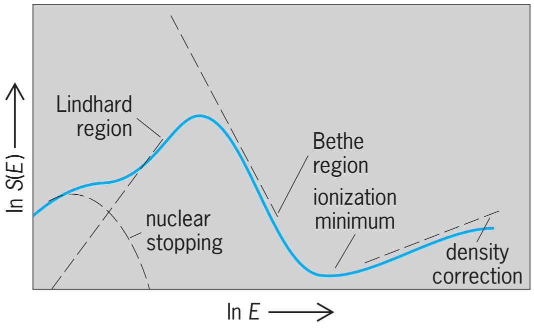


Fig. 2. **Graph of the logarithm of the stopping power *S*(*E*) of a heavy ion in matter as a function of the logarithm of its energy *E*.**

For energies above the ionization minimum the stopping power increases, but at substantially higher energies increases less rapidly. This flattening is known as the density effect and is the result of the rapid rearrangement or polarization of the electrons in the medium by the increasingly strong transverse electric field of the projectile. This rearrangement partially shields the projectile and so slows the rate of increase of the stopping power, but the suddenness of the motion results in the emission of electromagnetic radiation. When the speed of the projectile exceeds the speed of light in the medium, this radiation is called Cerenkov radiation, and its direction is determined by the velocity of the projectile and the medium's index of refraction. *See* CERENKOV RADIATION; RELATIVITY.

As a projectile's energy is lowered below the Bethe region, it begins to capture electrons from the medium. Very roughly, a new electron is captured when the velocity of the projectile in the medium is comparable to the velocity of the electron in its quantum-mechanical orbit about the projectile. The presence of these captured electrons surrounding the penetrating nucleus shields it from the electrons of the medium and thus reduces the stopping power, resulting in a peak in the curve (Fig. 2). However, collisions with target electrons continue to dominate the stopping power below this peak. This portion of the graph is labeled the Lindhard region and has a slope proportional to $E^{1/2}$.

For projectiles with energies typically below a few hundred keV, the collisions are not simply with electrons but are elastic and inelastic collisions between whole atoms. This is called nuclear stopping since the nuclei of struck atoms acquire significant amounts of kinetic energy.

The stopping mechanisms for ions in solids result in their having well-defined ranges. This behavior is distinctly different from the penetration of photons (light, x-rays, and so forth) in matter. Photon intensity in matter decreases exponentially with depth with a characteristic length, but individual photons at any depth have the same energy as they had at the surface. Thus, photons are destroyed but do not lose energy, while ions lose energy but are not destroyed.

In crystalline solids it is possible to pick certain directions where the electron density is significantly lower. Ions penetrating in these channels between crystal rows or planes experience fewer collisions than ions penetrating in other directions. In fact, because an ion's potential energy increases in all transverse directions away from the axis of such a channel, the ion will tend to be constrained to move along it. This phenomenon, known as channeling, has been studied extensively. At higher energies the ion can be excited to higher quantum-mechanical states through its resonant coherent interaction with the channel. When the ion relaxes, it emits radiation. *See* CHANNELING IN SOLIDS.

Radiation effects. The dissipation of a projectile's kinetic energy in a solid can result in a number of secondary phenomena. Near the surface, electrons given sufficient energy may directly escape. These are often accompanied by Auger electrons and x-rays emitted by atoms which have lost inner-shell electrons through collisions. *See* AUGER EFFECT.

In many insulating solids (including mica, glasses, and some plastics) the passage of an ion with a large electronic stopping power creates a unique form of radiation damage known as a nuclear track. The track is actually a cylindrical volume surrounding the trajectory of a single ionizing projectile which has been rendered much more susceptible to chemical attack. As a result, when the substance is chemically etched, conical pits visible under an ordinary microscope are produced where ionizing particles have penetrated. This is particularly significant since the passage of single projectiles is registered and may be subsequently observed. Although the mechanism of track formation is still the subject of theoretical research, track-forming materials are widely used to detect ionizing radiation. *See* PARTICLE TRACK ETCHING.

When the projectiles have energies in the nuclear stopping range, a number of additional phenomena may occur. Of particular significance is the formation of a collision cascade. In the nuclear stopping region it is relatively likely that the projectile will transfer significant amounts of energy to individual target atoms. These atoms will subsequently strike others, and eventually a large number of atoms within the solid will be set in motion. This disturbance is known as a collision cascade. Collision cascades may cause permanent damage to materials, induce mixing of layers in the vicinity of interfaces, or cause sputtering if they occur near surfaces. The dynamical description of collision cascades is extremely complicated because of the large numbers of particles involved. Computer modeling and a branch of statistical mechanics called transport theory have been most successful, although both have limitations. The comprehensive mathematical description of collision cascades with arbitrary energy densities, realistic interatomic potentials, and consistent treatment of target surfaces is an active area of research. *See* BOLTZMANN TRANSPORT EQUATION; CRYSTAL DEFECTS; ION BEAM MIXING; RADIATION DAMAGE TO MATERIALS; SIMULATION; SPUTTERING.

Although it is a much less ubiquitous phenomenon than collisional sputtering, particle emission from insulating surfaces can also be initiated by ions with energies in the electronic stopping region by mechanisms which are collectively termed desorption. Where sputtering is primarily a statistical process understandable in terms of classical mechanics, desorption is inherently quantum mechanical, involving specific electronic transitions. As a result, desorption can be initiated equally efficiently by ions, electrons, and photons. Very large and fragile biological molecules can be removed from surfaces virtually intact following the impacts of MeV heavy ions or pulsed laser irradiation. This has made it possible to mass-analyze a large number of biological molecules which could not previously be studied by conventional mass spectrometry because they were too fragile to be thermally evaporated. *See* DESORPTION.

Surface interactions. Much work in the field of ion-beam interactions with matter has dealt with effects to the material and the projectile which occur in the region of the surface. Projectiles with energies ranging from chemical bond energies to penetration energies can be used to study surface structure, prepare unique deposition layers, include unusual chemical reactions, and even select specific reaction mechanisms. For example, bombardment of a surface by 150 keV C^+ ions has been shown to be one route to the formation of diamondlike carbon. Similarly, by studying the atomic state of hyperthermal ions such as Li^{+} which scatter from the surface, information about both the structure of the surface and the nature of the interaction itself can be inferred, thus providing valuable additional insights to those which have been available previously from beam-foil spectroscopy. *See* BEAM-FOIL SPECTROSCOPY.

Complex and fascinating phenomena have been observed in low-energy ion-surface collisions. They include instances where ions appear to make multiple impacts in much the same way as a stone skips on the water. The electron-cyclotron-resonance (ECR) ion source has made it possible to study the impacts of extremely slow, yet very highly charged ions with surfaces. One result of this work is the recognition that at distances greater than 1 nm from the surface these nearly bare nuclei may capture significant numbers of electrons into their high-lying quantum states while lower-lying states may remain essentially empty. The resulting structures, called hollow atoms, may survive in this condition briefly after impact with the surface. *See* ION SOURCES.

Advances in computing have made it possible to simulate ion-surface impacts with far-reaching consequences for the development of intuition about these events (**Fig. 3**). Such simulations represent essential tools in the design and interpretation of experiments. *See* FULLERENE; MOLECULAR MECHANICS.

Applications. The most ambitious use of ion beams in industry is in the manufacture of integrated circuits. Ion implantation can be used to modify the electrical properties of semiconductors far more precisely than other techniques. As a result, much more elaborate structures can be produced. Ion implantation is an essential element of the manufacture of virtually all digital integrated circuits in use. *See* INTEGRATED CIRCUITS; ION IMPLANTATION.

Another commercial application of ion beams with significant promise is for materials modification. It has been demonstrated that ion implantation can greatly improve the wear and corrosion resistance of metals. One area that is expected to benefit from these advances is the technology of human prosthetic implants. *See* PROSTHESIS.

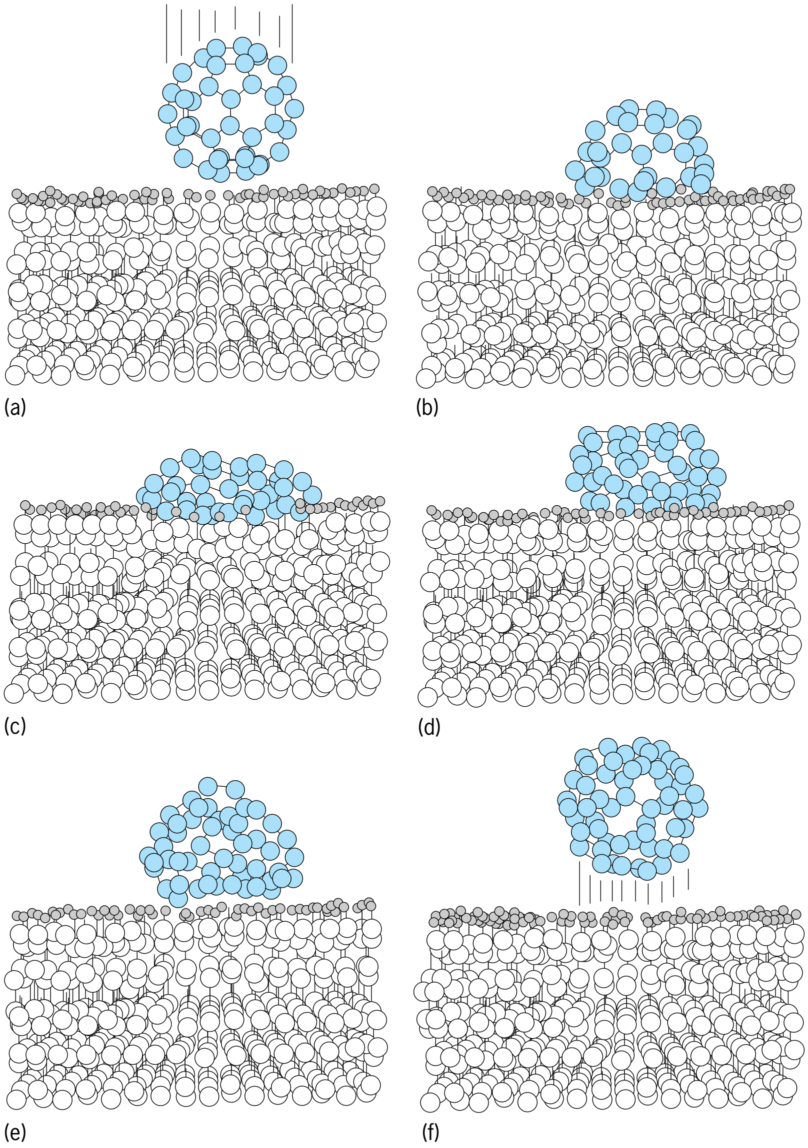


Fig. 3. **Molecular dynamics simulation of a 250-eV C_{60}^{+} molecular ion (buckminsterfullerene or buckyball) scattering from a hydrogen-terminated diamond surface. Atomic positions are shown at various times. (*a*) 0 femtosecond (*b*) 40 fs. (*c*) 100 fs. (*d*) 160 fs. (e) 201 fs. (*f*) 280 fs. (*After R. C. Mowrey et al., Simulations of C_{60} collisions with a hydrogen-terminated diamond [111] surface, J. Phys. Chem., 95:7138–7142, 1991*)**

Ion-solid processes are used extensively as tools in many areas of research. They permit highly sensitive analyses for trace elements, the characterization of materials and surfaces, and the detection of ionizing radiation. Techniques employing them include secondary ion mass spectrometry (SIMS) for elemental analysis and imaging of surfaces, proton-induced x-ray emission (PIXE), ion-scattering spectrometry (ISS), medium-energy and Rutherford backscattering spectrometries (MEBS and RBS), and nuclear reaction analysis (NRA) for elemental and isotopic depth profiling. Ion-solid interactions are also fundamental to the operation of silicon surface-barrier detectors which are used for the measurement of particle radiation, and of nuclear track detectors which have been used in research as diverse as the dating of meteorites and the search for magnetic monopoles. Among the notable products of ion-solid research are the development of ion-beam assisted deposition (IBAD) for the rapid production of thick protective coatings, and the demonstration of thin-film waveguides and waveguide lasers produced in insulators by ion implantation. *See* ACTIVATION ANALYSIS; CHARGED PARTICLE BEAMS; MAGNETIC MONOPOLES; PARTICLE DETECTOR; PROTON-INDUCED X-RAY EMISSION (PIXE); SECONDARY ION MASS SPECTROMETRY (SIMS); SURFACE PHYSICS.

Robert A. Weller

Bibliography. R. Behrisch (ed.), *Sputtering by Particle Bombardment*, 3 vols., 1981, 1983, 1991; W. K. Chu, J. W. Mayer, and M. A. Nicolet, *Backscattering Spectrometry*, 1978; J. P. Eberhart, *Structural and Chemical Analysis of Materials*, 1991; L. C. Feldman and J. W. Mayer, *Fundamentals of Surface and Thin Film Analysis*, 1986; J. R. Tesmer et al. (eds.), *Handbook of Modern Ion Beam Materials Analysis*, 1995; J. F. Ziegler, J. P. Biersack, and V. Littmark, *The Stopping and Range of Ions in Solids*, 1985.

Ion sources

Devices that produce positive or negative electrically charged atoms or molecules. *See* ION.

In general, ion sources fall into three major categories: those designed for positive-ion generation, those for negative-ion generation, and a highly specialized type of source designed to produce a polarized ion beam. The positive-ion source category may further be subdivided into sources specifically designed to generate singly charged ions and those designed to produce very highly charged ions.

Desirable qualities of an ion source are large yield, high ionization efficiency, low energy spread, high brightness, and low emittance. Practical considerations such as reliability, long source life, and ease of changing ion species are also important.

Ion sources have acquired a wide variety of applications. They are used in a variety of different types of accelerators for nuclear research; have application in the field of fusion research; and are used for ion implantation, in isotope separators, in ion microprobes, as a means of rocket propulsion, in mass spectrometers, and for ion milling. *See* ION IMPLANTATION; ION PROPULSION; ISOTOPE SEPARATION; MASS SPECTROSCOPE; NUCLEAR FUSION; PARTICLE ACCELERATOR; SECONDARY ION MASS SPECTROMETRY (SIMS).

Methods of Positive-Ion Formation

The principal methods of positive-ion formation are electron impact, surface ionization, spark discharge, laser ionization, field ionization, thermal ionization, and sputtering.

Electron impact. A common method of ionizing a gas or vapor is to pass high-velocity electrons through it, with the ions being formed as a result of electron-atom collisions. Electron energies are typically a few 100 eV but in some special sources may be as high as 20 keV. An externally applied magnetic field is frequently used to cause the electrons to travel along a helical path and thereby increase the ionization efficiency. Examples of ion sources utilizing this concept are the duoplasmatron and the Penning ion source.

Positive surface ionization. Atoms possessing low ionization potentials can be ionized by allowing them to strike a heated surface having a high work function. Provided that the ionization potential of the atom is less than or about equal to the work function of the surface, there is a high probability that the atom will be thermally desorbed as a positive ion. The method is particularly well suited, though not entirely restricted, to producing ions of the alkali metals, all of which have ionization potentials of less than 5.4 eV. Some high work function metals are platinum (approximately 5.7 eV), tungsten (approximately 4.5–5.5 eV), and rhenium (approximately 5 eV). *See* IONIZATION POTENTIAL; WORK FUNCTION (ELECTRONICS).

Spark discharge. There are several variations of this technique, but basically a spark is induced between two electrodes, one of which, at least, contains the element to be ionized. Generally speaking, the spark consists of a high-density, high-temperature plasma from which ions can be extracted. The spark can be produced by applying a high alternating potential between two fixed electrodes or by mechanically breaking contacting electrodes. *See* ELECTRIC SPARK.

Laser ionization. A focused beam from a high-power pulsed laser can be used to produce a small ball of dense plasma from essentially any solid, and positive ions can be extracted from this plasma. The high temperature of the plasma results in the formation of many multiply stripped ions and thus may prove a very effective method of generating highly charged positive ions.

In principle, lasers or other strong sources of electromagnetic radiation can be used to produce ions by photoionization. A photon can directly ionize an atom if its energy exceeds the ionization potential of the atom. Unfortunately, the probability of photoionization is low (the cross section is of the order of 10^{-19} cm^2), making it difficult to design efficient ion sources based on this process. *See* LASER.

Field ionization. If an atom passes close to or gets absorbed on a very sharp point where the electric field exceeds a few times 10^{10} V/m, there is a probability that it will get ionized; the phenomenon is known as field ionization. Such large electric fields can be achieved in the vicinity of a specially sharpened tungsten needle placed close to an annular electrode, and gas or vapor passing close to the tip of the needle can be ionized.

Field emission or ionization is generally believed to be the underlying operating principle of a novel type of ion source known as the electrohydrodynamic source. In this source a conducting liquid, usually a metal, is allowed to be drawn down a fine-bore tube by capillary action. When an electric field is applied to the tip of the tube, the liquid meniscus, normally spherical, distorts and becomes conical. As the electric field is increased, the tip of the cone becomes sharper, and eventually the field at the tip becomes sufficiently large to cause field emission. *See* FIELD EMISSION.

Thermal ionization. Although the term thermal ionization is ill-defined, it is generally used in the context of heating certain complex compounds, resulting in positive-ion emission. An example is the emission of lithium ions from a heated surface coated with β-eucryptite (a lithium aluminosilicate mineral). The technique has found extensive application in mass spectroscopy to produce a wide variety of ions. The sample to be ionized is usually mixed with silica gel and phosphoric acid and deposited on a rhenium filament. After a preliminary baking, the filament is introduced into the ion source and heated to the point of positive-ion emission.

Sputtering. When a solid is bombarded with energetic heavy ions, a fraction of the sputtered particles leaves the surface as ions. This fraction is usually too low for direct application in an ion source, but the sputtering process is frequently used to introduce solids into an electron impact source such as a Penning source. *See* SPUTTERING.

Methods of Negative-Ion Formation

All elements can be positively ionized, but not all form stable negative ions. For example, none of the noble gases forms negative ions. However, helium is an exception in that it does have a metastable negative ion with a lifetime of about 1 millisecond. The noble gases are not the only elements that do not form stable negative ions, but most form metastable ones with lifetimes long enough to permit acceleration. Nitrogen has an exceptionally short-lived metastable negative ion, and it is customary to accelerate either NH^- or NH_2^- molecular ions, both of which are stable.

Direct extraction. Most discharge sources, such as the duoplasmatron (**Fig. 1**), yield negative ions when the polarity of the extraction voltage is reversed. However, the yield is usually low, the electron current is high, and there are difficulties when operating them with elements other than hydrogen. These sources are now used almost exclusively to generate intense beams of hydrogen and its isotopes, and several important changes have been made to improve their performance. The negative-ion-to-electron yield from a direct extraction duoplasmatron can be greatly improved if the intermediate electrode is displaced by a millimeter or so off the source axis (Fig. 1*b*). The introduction of cesium vapor into the discharge plasma greatly enhances the negative-ion yield. Negative hydrogen currents from such sources have been increased from a few microamperes to several tens of milliamperes.

Charge exchange. When a positive-ion beam, ranging in energy from a fraction of a kiloelectronvolt to several tens of kiloelectronvolts, is passed through a gas or vapor, some of the ions emerge negatively charged. At the optimum energy (depending upon the ion), and with vapors of lithium, sodium, and magnesium, the negatively charged fraction is quite high, ranging from 1 to 90%. The technique is also highly effective for the creation of metastable negative ions, such as helium.

Cesium-beam sputtering. When a solid surface is sputtered with cesium positive ions, a surprisingly large fraction of the sputtered particles emerge as negative ions. This fraction can be increased by almost an order of magnitude by overlaying the sputter surface with additional neutral cesium. Highly efficient and versatile negative-ion sources involving cesium sputtering have been developed.

Negative surface ionization. Just as positive ions of elements possessing low ionization potentials can be generated by thermally desorbing them from high work-function surfaces, negative ions having large electron affinities can be similarly generated on a low work-function surface. The method is particularly suited to generating negative ions of the halogens, all of which have electron affinities in excess of 3 eV. A particularly suitable low-work-function surface is lanthanum hexaboride, which is reported to have a work function of about 2.6 eV.

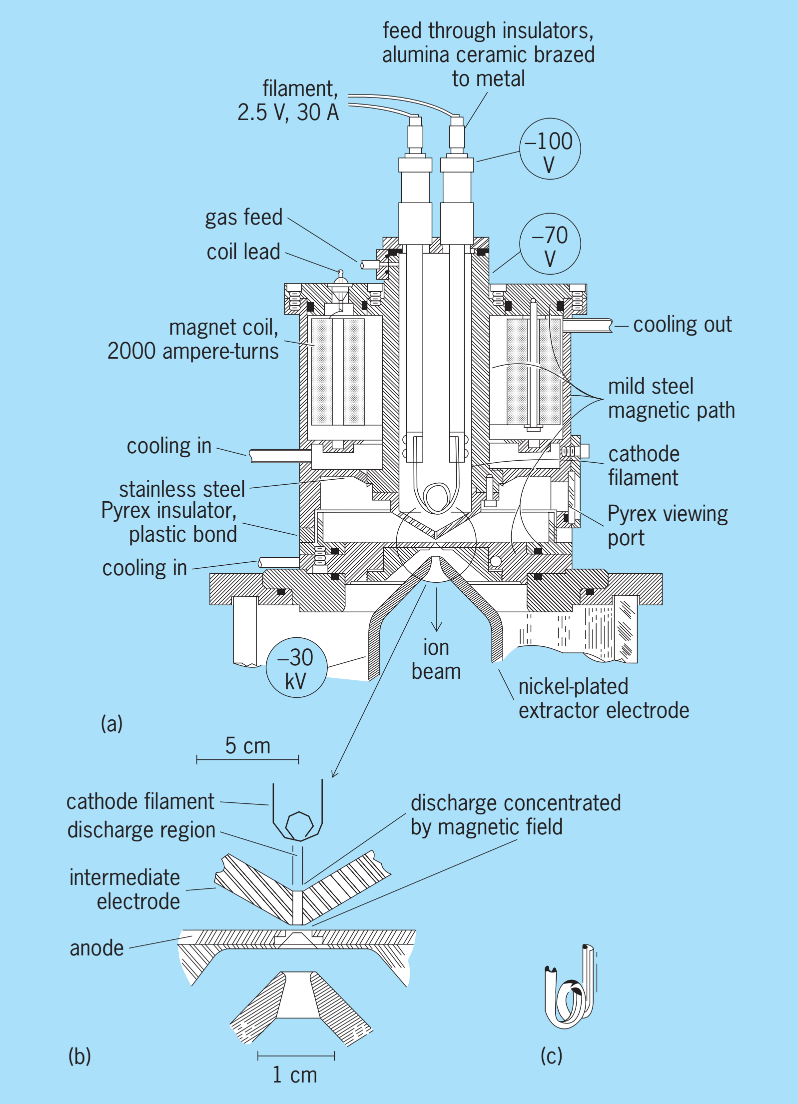


Fig. 1. Duoplasmatron ion source for accelerators. (*a*) Diagram of entire source. (*b*) Detail of discharge and extraction regions. (*c*) Drawing of cathode filament. 1 cm = 0.4 in.

Positive-Ion Source Concepts

Positive-ion source concepts include the duoplasmatron source for protons and the Penning ion source and ion confinement sources for multiply charged heavy ions.

Duoplasmatron. The duoplasmatron (Fig. 1) is a high-current proton source. It makes use of an arc discharge which is constricted as it passes into a very strong magnetic field shaped by iron or mild steel inserts in an intermediate electrode and anode. The beam is extracted at the point where the arc has reached a very small diameter and a very high brilliance. Sources of this type have been developed for accelerators. *See* ARC DISCHARGE.

Heavy-ion sources. The term heavy ion is used to designate atoms or molecules of elements heavier than helium which have been ionized. As was mentioned earlier, ion sources that are used to generate such ions fall into two categories: those intended to form singly charged ions and those designed to produce multiply charged ions. A heavy ion can be singly ionized (one electron removed or added), can be fully stripped as in argon 18+, or can have any intermediate charge state.

Singly charged heavy ions are most frequently used in isotope separators, mass spectrographs, and ion implantation accelerators. These are much easier to generate than multiply charged ions, and frequently the experimenter has several source concepts to choose from, depending upon the application and the physical characteristics of the element to be ionized.

Penning ion sources. This source is based on a high-current gaseous discharge in a magnetic field with gas at a relatively low pressure (10^{-3} torr or 0.1 pascal). The source (**Fig. 2**) consists of a hollow anode chamber, cathodes at each end, a means for introducing the desired element (usually a gas), and electrodes for extracting the ions (not shown). The cathode may be heated to emit electrons, which then help to initiate the arc discharge current, creating the plasma in which the atoms are ionized. The discharge column between the cathodes (the plasma) consists of approximately equal numbers of low-energy electrons and positive ions. The electron density is much larger than can be accounted for by the primary electrons from the cathodes. The average energy of plasma electrons may range from a few volts to a few tens of volts. Electrons travel parallel to the magnetic field, are reflected from the opposite cathode, and make many traversals of the length of the hollow chamber. The electrons confined by the magnetic field and the cathode potential thereby have a high probability of making ionizing collisions with any gas present in the chamber.

TABLE 1. Penning heavy-ion source performance

Element	Charge state distributions (I^{Q+}/I^{1+})								
	1+	3+	4+	5+	6+	7+	8+	9+	12+
Argon	1	8.5	8	3	0.8	0.09	0.01		
Calcium	1	23	22	15	3	0.3	0.035		
Krypton	1	5	5.5	6	4.5	4	2.2	0.6	
Xenon	1	7	9	9	7.5	5.5	4	1.5	0.025

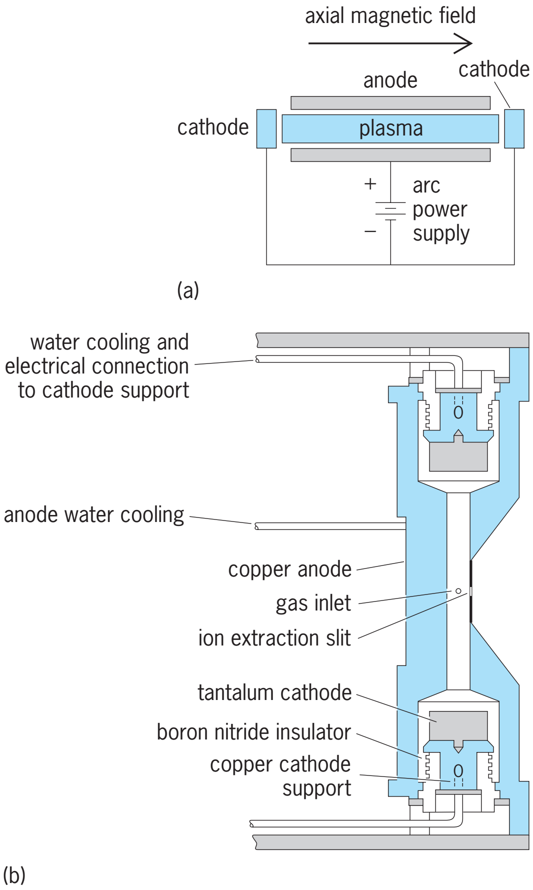


Fig. 2. Penning ion source. (*a*) Schematic diagram illustrating basic principles. (*b*) Section showing geometry. (*After J. R. J. Bennett, A review of PIG sources for multiply charged heavy ions, IEEE Trans. Nucl. Sci., NS-19*(2):*48–68, 1972*)

The net result of all the processes in the arc plasma is that some partially stripped atoms diffuse perpendicular to the magnetic field out of the arc, experience the field of the accelerating electrode, and are moved into the accelerator. The arc potential may be constant with time, or it may be pulsed so that ions are produced as needed by the accelerator. High yields of charge states 1+ through 8+ (and for heavier elements, perhaps up to 12+) have been obtained for many elements of the periodic table (**Table 1** shows four of these). *See* ARC DISCHARGE; PLASMA (PHYSICS).

Most studies with Penning sources were made with gases, thus restricting nuclear research programs to a rather limited number of projectiles, until it was discovered that some of the krypton ions from the source were partially accelerated and then returned back into the source where they sputtered the source body material (copper) into the discharge, resulting in a prolific beam of highly charged copper ions. It was a short step to introduce small pieces of various solids immediately behind the ion extraction slit (Fig. 2*b*) and to use the source to produce beams of a wide range of elements. Roy Middleton

Electron cyclotron resonance sources. An electron cyclotron resonance source for highly charged ions consists of a metallic box occupied by low-pressure vapors, microwaves, and specific magnetic fields for plasma confinement. The box can be of any shape but must be large with respect to the wavelength of the microwaves in order to act as a multimode cavity. Plasma confinement is generally obtained by superimposing solenoidal and multipolar fields such that the modulus of the magnetic field is minimal in the center of the box and maximal near the walls. In between, there should be a closed magnetic surface where the Larmor frequency of the electrons equals the frequency of the injected microwaves. Electrons crossing this surface are then energized by the electron cyclotron resonance. When they pass many times through the resonance, according to their random phase, they acquire a global stochastic electron cyclotron resonance heating, yielding

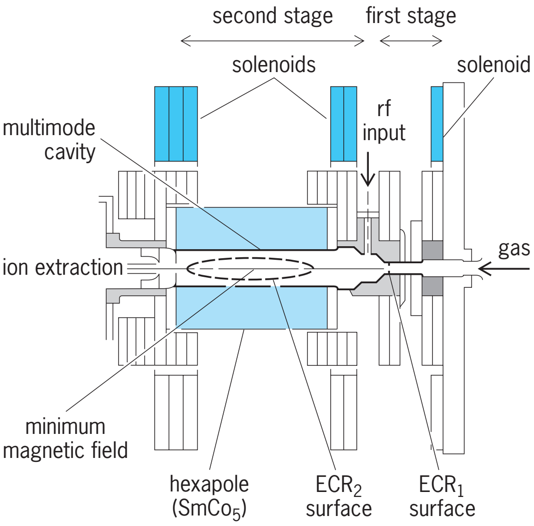


Fig. 3. Electron cyclotron resonance ion source.

energies of tens of kiloelectronvolts, thus exceeding the ionization potentials of many highly charged ions. However, since the ion charges are obtained step by step through successive electron–ion collisions, the ions need long exposure times to many electron impacts (that is, long plasma confinement). The source (**Fig. 3**) generally has two stages in series. In the first, a cold plasma is created at ECR_1, which diffuses toward a very low-pressure second stage, where ECR_2 energizes the electrons inside the confined plasma. Owing to the absence of electrodes, the longevity of the source is unlimited. The magnetic fields are obtained with ordinary or superconducting coils, permanent magnets,or combination of these. The operation of the source can be continuous or pulsed. Most sources utilize frequencies between 2.4 and 16 GHz. The performance improves with increasing frequency. *See* CYCLOTRON RESONANCE EXPERIMENTS; MAGNET; SUPERCONDUCTING DEVICES.

R. Geller

Electron-beam sources. The electron-beam ion source (EBIS) produces the most highly charged ions, up to fully stripped U^{92+}. It can be operated in many different modes depending on the application, the desired ion species, charge state, and ion-beam characteristics, including its duty cycle. The EBIS normally produces ion beams with a small emittance and a narrow energy spread. The highest charge states are normally produced in a batch mode, yielding ion pulses with a charge as high as 1 nC, which allows for instantaneous peak currents as high as 1 μA. The time-averaged currents are normally in the range of nanoamperes to picoamperes because of the required batch production time, which increases with the desired charge state.

The majority of EBISs or electron-beam ion traps (EBITs) are used to study how the produced highly charged, low-energy ions interact with the internal electron beam (in the EBIT) or with external gaseous or solid targets (in the EBIS). Others are used as injectors for pulsed accelerators, storage rings, or ion traps. A few are used to study further developments or new applications.

The ion source (**Fig. 4**) is based on an electron beam which successively strips the ions until the desired high charge state is reached. The electron beam is launched from an electron gun, confined and normally compressed by a strong magnetic solenoid, and later adsorbed by a collector. Ions trapped inside

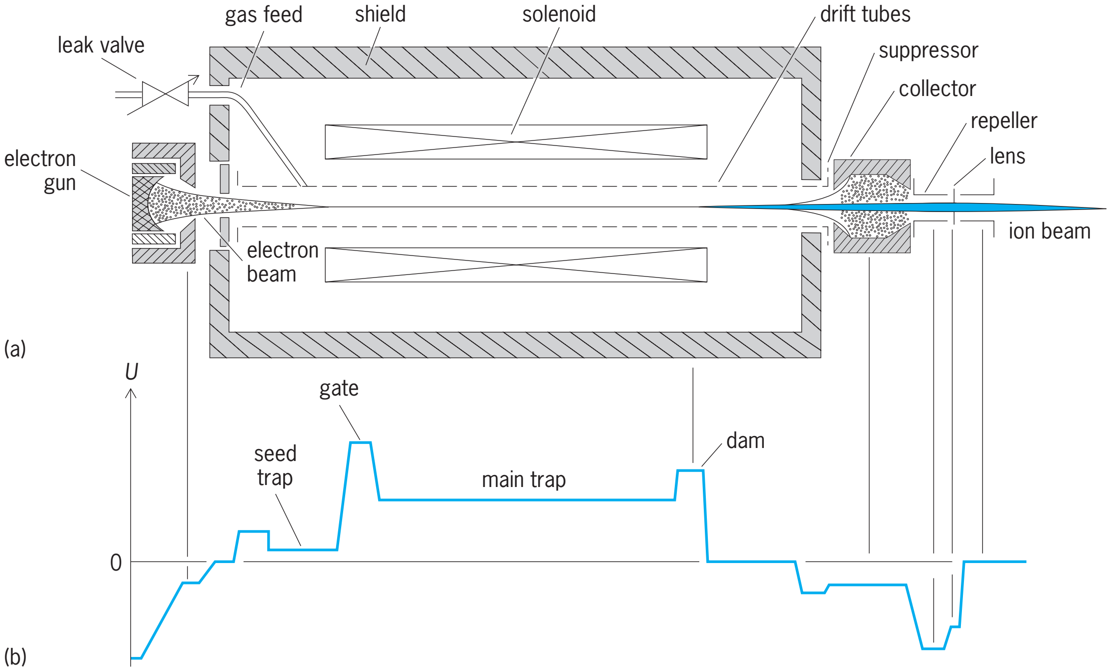


Fig. 4. Electron-beam ion source. (*a*) Cross section. (*b*) Electric potential (*U*) applied to the drift tubes during the production of highly charged ions.

the space charge of the electron beam are ionized at a rate which depends on the electron beam's energy and density, with high-power electron beams and high-field solenoids producing high rates. The charge states produced increase with increasing confinement time until the ionization rate is matched by the rate of electron capture mainly in collisions with the residual gas. Hence the highest charge states are produced with cryogenic EBISs which use a superconducting high-field solenoid (3–6 tesla) to achieve high electron-beam compression and use the required liquid helium to cryopump the production region to reduce the residual gas densities. The electron beam is surrounded by a series of drift tubes, with their potential being used to adjust the electron-beam energy and to control the flow of the ions, allowing for the variation of the confinement time. The highest charge states are normally produced in a batch mode which begins with the injection of low-charged seed ions either from an external ion source or from an internal seed trap in the fringing fields of the solenoid, where the seed gas is ionized by the lower-density electron beam. After the gate or the dam has been raised, the ions are confined in the axial direction with potentials that exceed the potential applied to the intermediate drift tubes, the main trap. After the confinement time required to reach the desired charge state (10^{-4} to 50 s), the main-trap potentials is raised until the ions escape the source. A slow rise of the main-trap potential decreases the ion beam's energy spread and simultaneously decreases the instantaneous current while increasing the duty cycle. *See* MAGNET; SOLENOID (ELECTRICITY); VACUUM PUMP. Martin P. Stockli

Negative-Ion Source Concepts

Negative-ion source concepts will be discussed, including the methods of negative-ion formation discussed above.

Charge exchange source. Although charge exchange can be used to provide a very large variety of negative ions, it is relatively infrequently used, with the exception of producing the metastable negative ion of helium. Indeed, charge exchange is the sole method of producing negative helium beams.

In a typical charge exchange source, about 1 milliampere of positive helium ions is generated in a duoplasmatron ion source which is usually at ground potential. The ions are extracted by an electrode at a potential of about −20 kV, focused by an electrostatic lens, and directed through a donor canal usually containing lithium vapor. In the vapor a fraction of the positively charged helium ions sequentially picks up two electrons to form negative ions. The fraction is usually a little less than 1%, resulting in a negative helium beam of several microamperes.

Cesium-beam sputter source. The cesium-beam sputter source is the negative-ion source most widely used on tandem accelerators. Much of its success is due to the fact that ion species can be changed rapidly, the ion output is usually high (**Table 2**), and the source will operate for several hundreds of hours before a major cleanup is necessary.

TABLE 2. Negative-ion currents from cesium-beam sputter source

Element	Negative-ion current, μA	Element	Negative-ion current, μA
Lithium	2	Sulfur	20
Boron	2	Nickel	6
Carbon	>50	Copper	3
Oxygen	>50	Gold	10
Silicon	20	Lead	0.1

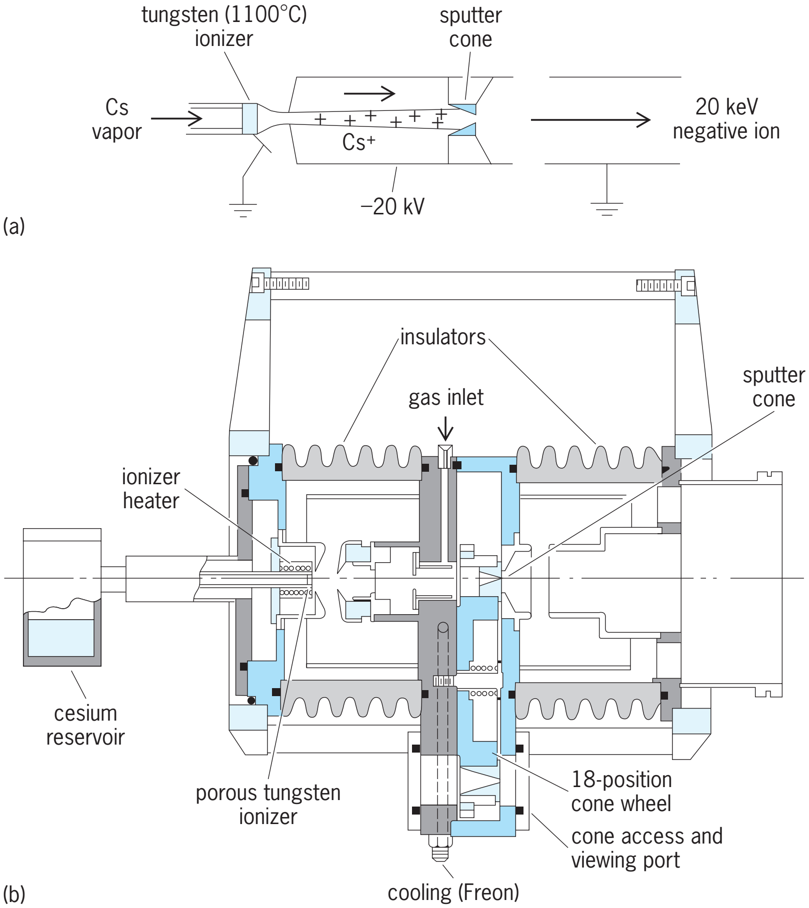


Fig. 5. Cesium-beam sputter source: (*a*) schematic showing operating principles; (*b*) section of a typical source.

Cesium positive ions are formed in the source (**Fig. 5*a***) by passing cesium vapor through a porous disk of tungsten, heated to about 2012°F (1100°C), by the surface ionization process. These ions are accelerated through a potential difference of between 20 and 30 kV and are allowed to impinge upon a hollow cone that is either fabricated from or contains the element whose negative ion is required. Sputtering of the cesium-coated inner surface of the cone results in a large fraction of the sputtered particles leaving the surface as negative ions. An appreciable fraction of these are extracted from the rear hole of the sputter cone because of electric field penetration and are accelerated toward the ground electrode as a beam.

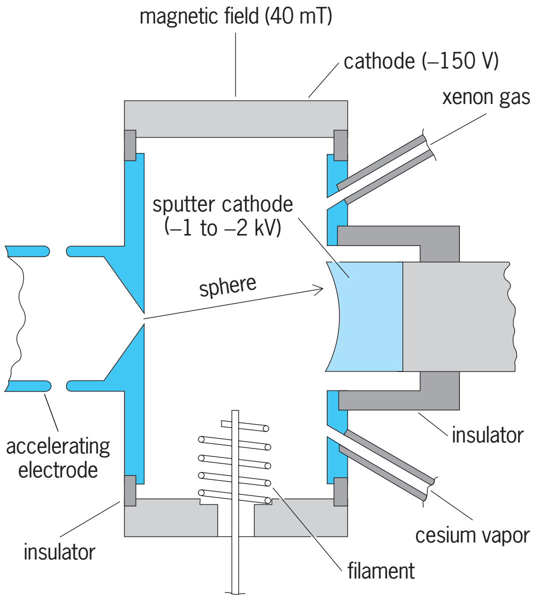


Fig. 6. Cesium-vapor Penning source for generation of negative ions.

The sputter cones or targets can be inserted into a cooled copper wheel resembling the chamber of a revolver (Fig. 5*b*). Thus, by rotating the wheel, sputter cones can be rapidly changed, enabling the negative-ion species to be quickly changed. In addition, negative-ion beams of gaseous elements, such as oxygen, can be formed by leaking the gas into the source and directing the flow onto a suitable getter surface such as titanium.

Cesium-vapor Penning source. The cesium-vapor Penning source (**Fig. 6**) is a direct-extraction negative-ion source. Basically this is a conventional Penning source but with two important modifications making it suitable for negative-ion generation. The first is the introduction of a third sputter cathode, which is the source of negative ions. This cathode is made from or contains the element of interest, has a spherical face centered on the extraction aperture, and is operated at a higher negative potential than the normal cathodes. The second change involves introduction of cesium vapor into the arc chamber in addition to the support gas (usually xenon).

The source operates in the normal Penning mode, and some of the cesium vapor introduced into the arc chamber becomes ionized and is accelerated toward the third sputter cathode. The negative ions that are formed as a result of sputtering are focused and accelerated toward the extraction aperture, and under the influence of the strong electric field generated by the acceleration electrode are extracted as an intense low-divergence beam. The negative-ion yield of the source is quite good and is comparable to that of the cesium-beam sputter source. Roy Middleton

Polarized Ion Sources

A polarized ion source is a device that generates ion beams in such a manner that the spins of the ions are aligned in some direction. The usual application is to inject polarized ions into a particle accelerator; however, it also has applications in atomic physics. The possible types of polarized sources are numerous, for in theory the nuclei of all kinds of atoms can be polarized, provided their spin is not zero. *See* SPIN (QUANTUM MECHANICS).

The original type of source, developed in 1956, generates only positive ions, while the metastable-state or Lamb-shift type of polarized ion source produces a high-quality negative ion beam with a high degree of polarization. The older type of source is referred to as the conventional or ground-state type of polarized ion source. Its output current of positive ions is an order of magnitude larger than the negative-ion output from the Lamb-shift source. With these two types of sources and their variants, polarized ions have been obtained from hydrogen, deuterium, tritium, helium-3, both isotopes of lithium, and others. The extra complication involved in producing polarized ions is such that the output is a factor of a thousand or more below the output of a moderately sized unpolarized ion source.

Conventional or ground-state source. In this type of source (**Fig. 7**), the first step consists in forming a beam of atoms in the ground state by a technique similar to that used in molecular beams. In the case of hydrogen or deuterium, this is done by dissociating the molecular gas in a glass discharge tube and allowing the atoms to escape through a nozzle. The atoms escaping at thermal energies are collimated

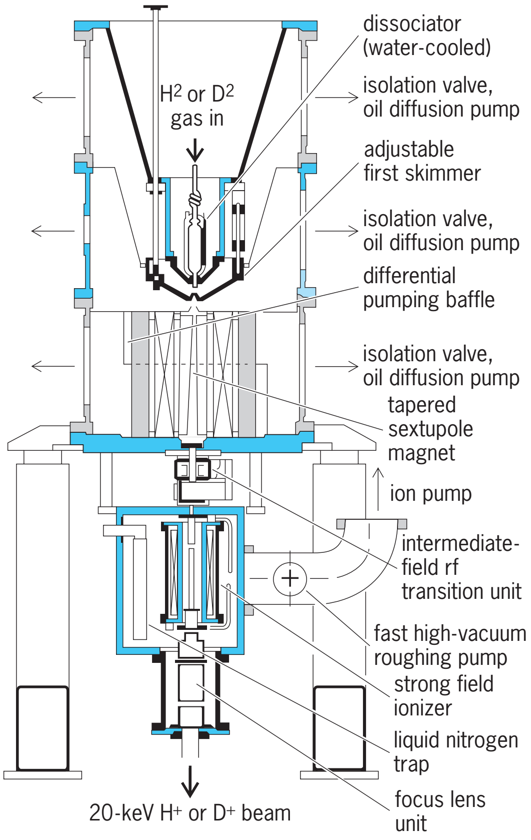


Fig. 7. Conventional or ground-state polarized ion source. (*After H. F. Glavish, Polarized ion sources, in Proceedings of the 2d Symposium on Ion Sources and Formation of Ion Beams, Berkeley, pp. IV-1-1 through IV-1-7, 1974*)

into a well-directed beam by plates with holes or an adjustable skimmer. High-capacity diffusion pumps sweep away the large quantity of excess hydrogen or deuterium. *See* MOLECULAR BEAMS.

The beam is then passed along the axis of an inhomogeneous magnetic field, which is most commonly generated by a sextupole magnet. This type of magnet consists of six magnets arranged in a circular pattern with alternating polarities.

In a sextupole magnet the absolute magnitude of the field increases as the distance squared from the axis. The atoms are subjected to a radial force that is proportional to their magnetic moment times the gradient of the absolute magnitude of the field strength. In the case of a sextupole magnet the force is proportional to the first power of the distance from the axis. The sign of the force does not depend upon the direction of the magnetic lines, but only on the projection of magnetic moment along these lines of force or m_j, where j is the spin value of the electron. (The atomic magnetic moment results almost entirely from the electron). The result is that atoms with a positive value of m_j are subjected to a force that is directed radially inward and pass through the sextupole magnet, while atoms with a negative m_j experience a force that is directed outward and are rapidly lost from the beam. Out of the sextupole comes a beam of atomic hydrogen that is polarized with respect to the orientation of its electrons but is, as yet, unpolarized in its nuclear spin. *See* ELECTRON SPIN; MAGNETON.

Since aligned nuclei rather than aligned electrons are desired, it is necessary to subject the atomic beam to other fields. Each hydrogen atom is in one of two pure states. It is possible to apply an oscillating magnetic field in combination with a dc magnetic field that will flip the sign of m_I (the projection of the nuclear spin) of one of the pure states and not the other. That aligns the spins of the nuclei but may depolarize the electrons. That does not matter, however, since they will be removed.

The final stage is to send the atomic beam into a strong solenoidal magnetic field. As the atoms from the sextupole field—having all orientations in each cross-sectional plane—enter the solenoid, they adiabatically come into alignment with the parallel lines of force within the solenoid since their m_j components of spin are conserved. In the solenoid the atoms are ionized by energetic electrons as in an arc discharge. The ionizer is actually the most difficult part of this type of polarized source to make function efficiently, even though it is conceptionally simple. The ionizer is followed by electric fields that accelerate and focus the ions to get a beam that can be accepted by the accelerator.

Lamb-shift or metastable-atom source. The polarization process in the Lamb-shift type of source is also performed upon atoms, in this case, metastable ones. The process is most efficient if the atoms have a velocity of approximately 10^{-3} of light rather than thermal velocity as in the case of the ground-state type of source. To get the beam, hydrogen, deuterium, or tritium can be used, but only hydrogen is discussed in this article. The hydrogen is ionized in a conventional ion source such as a duoplasmatron. The H^+ ions are then accelerated and focused into a beam at about 500 eV. The beam is passed through cesium vapor where cesium atoms donate electrons which are resonantly captured in an $n = 2$ state by the hydrogen ions.

Atoms are formed in both the $2p$ and the $2s$ states in the cesium vapor. However, those in the $2p$ state decay almost immediately to the ground state. The small energy difference between the $2p$ and the $2s$ states is the Lamb shift. The lifetime of the $2p$ atoms is 1.6×10^{-9} s, while the lifetime of the $2s$ atoms is 0.15 s because two photons must be emitted simultaneously in their decay to the ground state. Actually few $2s$ atoms decay by emission of two photons for they are necessarily subjected to small electric fields which mix into the $2s$ the $2p$ wave function and its tendency to decay to the ground state. To take advantage of this tendency to decay to the ground state, apparatus can be built so that those atoms which have the undesired value of m_I are stimulated to decay, while those with the desired value of m_I are allowed to pass on without decay. *See* ATOMIC STRUCTURE AND SPECTRA; FINE STRUCTURE (SPECTRAL LINES).

The polarized H^- ions are formed in argon because its atoms are capable of donating electrons to metastable atoms but have a very weak capability of forming H^- ions out of ground-state atoms. The ground-state charge-changing cross section appears to be lower by a factor of about 400; however, ground-state atoms outnumber the metastable atoms at this region by a factor of 40 so that the net polarization is 90%. The remainder of the apparatus consists of electric fields that accelerate and focus the beam so it can be accepted by an accelerator.

The electron spins are polarized by applying a transverse field of about 100 kV/m while the atomic beam of metastables is passing along the axis of a solenoid at a field strength of about 57.5 milliteslas. The transverse electric field couples the $2s$ and $2p$ levels through the Stark effect, and the magnetic field is just sufficient to bring the levels with $m_j = -{}^1\!/_2$ very close together in energy, while those with $m_j = +{}^1\!/_2$ have their energy separation doubled, so that $2s$ atoms with $m_j = +{}^1\!/_2$ are transmitted without loss. *See* STARK EFFECT; ZEEMAN EFFECT.

There are several methods of going on to polarize the nuclei, including a device known as the spin filter. To produce the spin filter, a longitudinal electric field of about the same strength as the transverse field is added to the apparatus that polarizes the electrons, with the longitudinal field oscillating at about 1.60 GHz. If the magnetic field is adjusted so the Larmor frequency of the electron in the metastable atom is made equal to the oscillating electric field, then the lifetime of the atom for decay becomes very long exactly at resonance, yet short not far off resonance. The magnetic field that determines the Larmor frequency of the electron in the metastable atom is the sum of that due to the solenoid and that due to the proton aligned in the solenoidal field. These two fields have opposite signs in the case of $m_I = -{}^1\!/_2$, and it is found that the two resonances for transmission are at 54.0 mT for $m_I = +1/2$ and at 60.5 mT for $m_I = -1/2$. In the case of deuterons, there are three resonances, and they are well resolved even though $m_I = +1$ is at 56.5 mT, 0 is at 57.5 mT, and -1 is at 58.5 mT. Joseph L. McKibben

Colliding-beam source. A novel method of making polarized beams of negative hydrogen and deuterium is based on the direct conversion of polarized neutral hydrogen (H^0) or deuterium (D^0) atoms into polarized negative ions using the reaction below.

$$H^0 \text{ (polarized)} + Cs^0 \rightarrow H^- \text{ (polarized)} + Cs^+$$

Although conceptually very attractive, this idea presents some severe experimental difficulties. Undoubtedly the greatest of these is that if the polarized H^0 or D^0 atoms are produced at thermal energies, where production is greatly facilitated, the cross section or probability of the above reaction proceeding is extremely small. To circumvent this difficulty and to capitalize on a much higher cross section, it was proposed to accelerate a positively charged cesium (Cs) beam to an energy of about 40 keV, neutralize it by passing it through a canal containing cesium vapor, and allow this high-velocity beam of neutral cesium atoms to collide with the polarized atomic beam.

Such a source has been built and demonstrated to yield 2.9 microamperes of polarized hydrogen negative ions and 3.1 μA of deuterium ions. These currents are about five times larger than those obtainable from the best Lamb-shift sources. Roy Middleton

Bibliography. L. W. Anderson (ed.), *Polarized Ion Sources and Polarized Gas Targets*, 1994; I. G. Brown (ed.), *The Physics and Technology of Ion Sources*, 1989; A. Hershcoritz (ed.), *International Symposiumon Electron Beam Ion Sources and Their Applications*, 1989; R. Marrus (ed.), *Physics of High-Ionized Atoms*, 1989; J. Parker (ed.), *International Conference on ECR Ion Sources*, Michigan State University, C.P.47, 1987; G. Roy and P. Schmor (eds.), *Polarized Proton Ion Sources, Conference Proceedings*, TRIUMF, Vancouver, 1983.

Ion transport

Movement of salts and other electrolytes in the form of ions from place to place within living systems.

Animals

Ion transport may occur by any of several different mechanisms: electrochemical diffusion, active-transport requiring energy, or bulk flow as in the flow of blood in the circulatory system of animals or the transpiration stream in the xylem tissue of plants.

Sodium/potassium pump. The best-known system for transporting ions actively is the sodium/potassium (Na/K) exchange pump which occurs in plasma membranes of virtually all cells. The pump is especially concentrated in the cells of membranes such as frog skin, the tubules of vertebrate kidneys, salt glands in certain marine reptiles and birds, and the gills of fresh-water fishes and crustaceans, which transport sodium ion (Na^+) actively.

The physical basis of the Na/K pump has been identified as a form of the enzyme adenosine triphosphatase (ATPase) which requires both Na^+ and K^+ for its activity of catalyzing the hydrolysis of adenosine triphosphate (ATP). This reaction provides the energy required for the extrusion of Na^+ from cells in exchange for K^+ taken in, but the mechanism of coupling of the energy liberation with the transport process has not been established. The ATP which provides energy for this and other processes of cellular work is formed by synthesis coupled with cellular oxidations. Again, the precise coupling mechanism linking oxidation with synthesis of ATP has not been established heretofore.

Chemiosmotic hypothesis. The basis of the hypothesis was the concept that, if the enzyme system responsible for the transport process is embedded in the plasma membrane with an orientation such that the direction of the reaction catalyzed is parallel to the directional or vectorial orientation of the transport across the membrane, then the vectorial character of transport processes may be explained.

Proton pumps. Although this hypothesis was not generally accepted, experimental studies revealed that many transport processes, such as in bacterial cells and in the mitochondria of eukaryotic cells, are associated with a transport of protons (hydrogen ions, H^+). This fact led to the concept of proton pumps, in which the coupling or transfer of energy between oxidation processes and synthesis of ATP and between hydrolysis of ATP and transport or other cellular work is explained in terms of a flow of protons as the means of energy transfer.

Electron transport system. The processes of oxidation in the citric acid cycle of reactions in mitochondria are known to be coupled with the synthesis of ATP, which is formed from adenosine diphosphate (ADP) and inorganic orthophosphate (P_i), through the system of enzymes and cytochromes known as the electron transfer chain or electron transport system. This system transports electrons, removed in dehydrogenation from the organic molecules of the citric acid cycle on one side of the mitochondrial membrane, to the site of their incorporation into water, formed from two hydrogen ions and an atom of oxygen on the other side of the membrane (**Fig. 1**). The flow of electrons from a relatively high potential level in the organic substrate to a level of lower potential in water constitutes, in effect, a current of negative electricity, and it was proposed that the flow drives a flow of protons in the opposite direction, as a current of positive electricity. This proton flow in turn is proposed as the force that drives the synthesis of three molecules of ATP for every two electrons flowing through the electron transport system. In effect, this is the machinery of the cellular power plant.

Na/K ATPase pump. The Na/K ATPase pump (**Fig. 2**) then provides an example of a way in which a proton pump may transfer energy between the

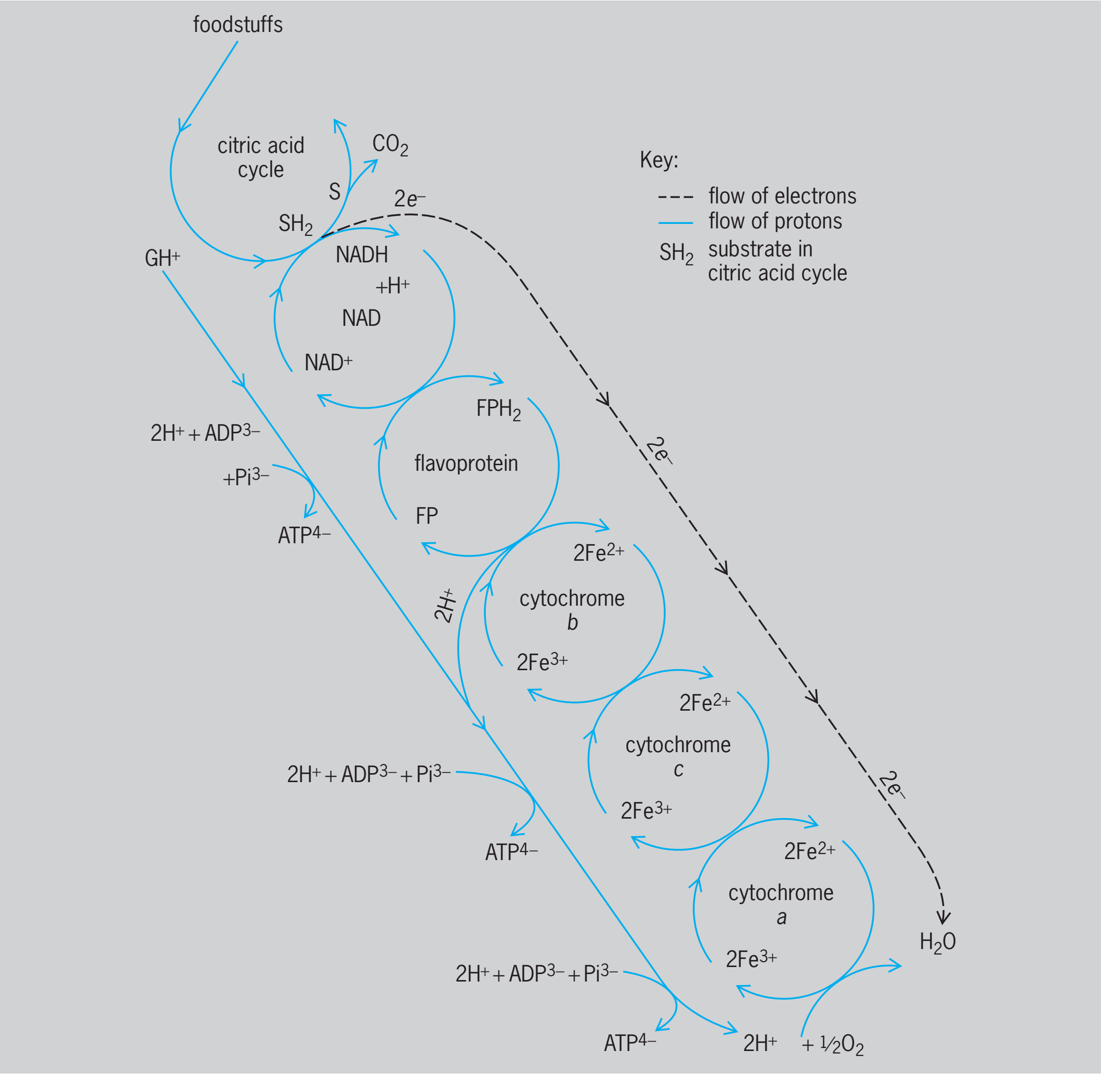


Fig. 1. Electron transport system as a proton pump.

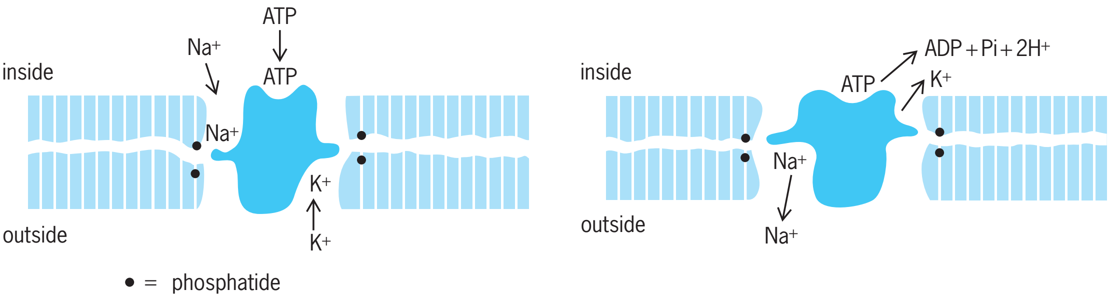


Fig. 2. Hypothetical model for the Na/K ATPase pump.

hydrolysis of ATP and a process of cellular work. The enzyme which is the basis of the pump is known to be bound to the lipid bilayer of the plasma membrane through phosphatides and to function only when so bound. The binding of Na$^+$, K$^+$, H$^+$, and ATP to active sites on the enzyme presumably has an allosteric effect, changing the shape of the enzyme molecule, activating the hydrolysis of ATP, and opening pathways of exchange of Na$^+$ and K$^+$. If this exchange were to be accompanied by a flow of H$^+$ as suggested in Fig. 2, the pump would be electrogenic, contributing to the internal negativity characteristic of nearly all cells by a net outward transport of positive electricity.

Bradley T. Scheer

Plants

Transport processes are involved in uptake and release of inorganic ions by plants and in distribution of ions within plants, and thus determine ionic relations of plants.

Scope. Photoautotrophic plants require nutrition with inorganic ions for growth and maintenance, that is, for osmoregulation, for sustainment of structure, and for control of metabolism. Ion transport physiology is concerned with the mechanisms of movement along various pathways of transport on organizational levels differing in their degree of complexity. Furthermore, ion transport is the basic functional component of some physiological reactions in plants. In several cases, ion transport processes are the physiological basis for ecological adaptations of plants to particular environments.

Organization levels. In unicellular, filamentous, or simple thalloid algae, in mosses, in poorly differentiated aquatic higher plants, and experimentally in cell suspension cultures of higher plants, ion transport can be considered on the cellular level. In these systems, all cells take up ions directly from the external medium. The cell wall and the external lipid-protein membrane (plasmalemma) have to be passed by the ions. Intracellular distribution and compartmentation are determined by transport across other membranes within the cells. The most important one is the tonoplast separating the cell vacuole from the cytoplasm.

Within tissues the continuous cell walls of adjacent cells form an apoplastic pathway for ion transport. A symplastic pathway is constituted by the cytoplasm extending from cell to cell via small channels of about 40 nanometers diameter (plasmodesmata) crossing the cell walls. Transport over longer distances is important in organs (roots, shoots, leaves, fruits), which are composed of different kinds of tissues, and in the whole plant. Xylem and phloem serve as pathways for long-distance transport. Roots take up ions from the soil and must supply other plant organs. But there is also circulation within the plant; for example, potassium, an inorganic cation of major importance in plants, is readily transported both from root to shoot and in the opposite direction. A cooperation of roots and shoots is observed during reduction of sulfur and nitrogen. The major nutrient elements sulfur, nitrogen, and phosphorus are taken up in the form of the inorganic anions, sulfate, nitrate, and phosphate. Phosphate is used in metabolism in the oxidized form in which it is available to the plant. NO_3^- and SO_4^{2-} must be reduced. The reduction capacity of roots can be limited. The anions are transported via the xylem to the leaves, where reduction proceeds using photosynthetic energy, reduction equivalents, and carbon skeletons for the attachment of reduced nitrogen and sulfur. Reduced nitrogen and sulfur compounds can be exported via the phloem. The nutritional status of roots and shoots regarding both inorganic anions and organic substrates plays a large role in regulation of ionic relations of whole plants. Phytohormones affect transport mechanisms; they are produced in particular tissues, are distributed via the transport pathways, and thus exert a signaling function. *See* PHLOEM; PLANT HORMONES; PLANT TISSUE SYSTEMS; XYLEM.

Transport mechanisms. There are a variety of basic and special mechanisms for the different pathways of ion transport.

Apoplastic and symplastic transport. The cell wall contains amorphic hemicellulosic and pectic substances and proteins. These constituents provide fixed electrical charges for ion exchange, which are largely negative because of the carboxyl groups of pectic acids; positive charges arise from the amino groups of proteins. The cellulose micelles and fibrils of the cell wall have free spaces of up to 10 nm diameter between them, allowing transport in an aqueous phase. Thus transport in the cell wall, that is, the apoplastic space, is based on physical mechanisms and driving forces such as ion exchange, diffusion in solution, and flow of solution.

After uptake via the plasmalemma, ion transport in the symplast is also maintained by physical driving forces. But metabolic energy is required for maintenance of the transport pathway and of concentration gradients. Cytoplasmic streaming can also play a role. *See* CELL WALLS (PLANT).

Membrane transport. A major issue in ion transport is movement across lipid-protein membranes. It is governed by physical and metabolism-dependent driving forces. Diffusion of ions leads to a gradient of electrical potential and of ion concentration at the membrane.

In plant cells, several mechanisms of primary active transport use metabolic energy (ATP or pyrophosphate, PP_i) directly to pump protons or calcium ions across membranes. The proton pumps establish electrochemical gradients of protons across the membranes. The calcium pumps participate in keeping cytosolic Ca^{2+} concentrations at a low level, which is important for the role of Ca^{2+} as a secondary messenger regulating cellular functions. The proton gradients serve secondary active transport since they can drive transport of other ions and of nonelectrolyte solutes. In this way, most ion transport processes at membranes of plants can be explained. Catalyzers of transport or carriers are involved. Transport often shows saturation kinetics with increasing concentration of the transported solute.

Transport in organs. Several mechanisms, corresponding to the various pathways, are involved in transport in organs. The most important example are roots (**Fig. 3**), although similar evaluations can be made for leaves or for specialized systems such as salt glands. Ions are absorbed from the soil by ion exchange at the root surface or by diffusion. The apoplastic pathway is blocked at a cell layer somewhat inside the root (the endodermis) where the cell walls are encrusted with an amorphic hydrophobic polymer (Casparian strips of endodermal cells). Thus, ions must cross the plasmalemma and pass the cytoplasm on their way to the xylem in the interior of the roots. Depending on the relative sizes of the various driving forces and resistances, both the apoplastic and the symplastic pathways can be used up to the endodermis. Two membrane transport processes involving metabolism-dependent ion pumping are involved: uptake from the medium or apoplast into the symplast at the plasmalemma in the root tissue peripheral to the endodermis; transport from the root cells lining the xylem vessels into these vessels.

Long-distance transport in whole plant. The pipe system of the xylem in its mature transporting state is

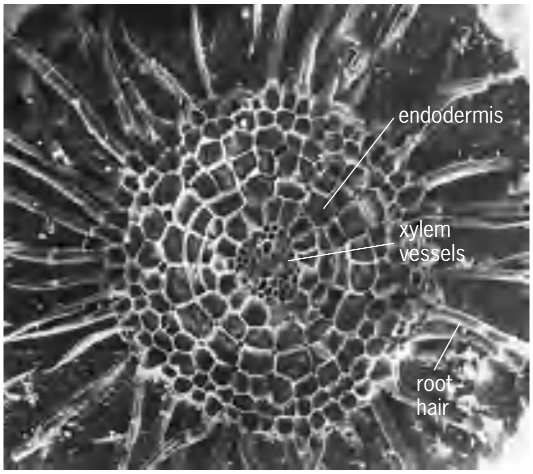


Fig. 3. Scanning electron micrograph of the cross section of a sunflower root, about 1 mm in diameter.

composed of rows of dead cells (tracheids, tracheary elements) whose cross-walls are perforated or removed entirely. The driving force for long-distance transport in the xylem is very largely passive. Transport is caused by transpiration, the loss of water from the aerial parts of the plant, driven by the water potential gradient directed from soil to roots, leaves, and atmosphere. A normally much smaller component driving the ascent of sap in the xylem is osmotic root pressure due to the pumping mechanisms concentrating ions in the root xylem, with water following passively. This can be demonstrated by attaching capillaries to the cut ends of isolated roots; it is the reason for the phenomenon of guttation, the exit of little droplets of sap from small openings in the leaves (**Fig. 4**). Root pressure amounts to only a few atmospheres, and, taking the resistance of the transport pathway into account, it can only supply shoots of small plants.

Fig. 4. Guttation droplets at the leaf dents of lady's-mantle (*Alchemilla*).

In a simplifying way the xylem can be considered as pathway for long-distance transport of ions from root to shoot, and the phloem for metabolite transport from photosynthesizing source leaves to various sinks in the plant. But inorganic ions can also move in the phloem, and this is important for recirculation in the plant. Only a limited number of ions are exempt from translocation in the phloem (such as Ca^{2+}). The long-distance transport pathways of the phloem are the sieve tubes, pipe systems with porous structures in the cross-walls (sieve plates) but, in contrast to vessels of the xylem, having living cytoplasm. Concentration and pressure gradients built up by active loading and unloading of sieve tubes in the source and sink regions, respectively, are the driving forces for transport.

Special adaptations. Ion transport mechanisms can provide the physiological basis for the adaptation to various environmental stress situations. For instance, iron-efficient dicotyledonous plants can increase the electron transport capacity at the plasmalemma of the cells on their root surfaces for reduction of Fe^{3+} to Fe^{2+} and hence for improved uptake of iron. Calcicole plants live only on calcareous soils, calcifuge plants exclusively on acid low-Ca^{2+} soils. These preferences are genetically fixed. There are also genotypic adaptations to high levels of various metal ions such as aluminum, zinc, copper, nickel, and lead. Particular transport properties of plants are of agricultural importance and thus interesting for breeders. There is also responsiveness to changes in the environment. In response to iron deficiency, some plants can thicken their root tissue behind the tips and form infoldings of the walls of peripheral root cells, increasing the plasmalemma surface for facilitation of transport. These structures disappear again as stress is released.

The most important example of stress is salinity due to NaCl. Some plants are moderately resistant by excluding NaCl at the level of the roots. The uptake into the root has K^+/Na^+ selectivity; K^+ can be taken up preferentially by exchanging for Na^+, which is released from the root tissue. Na^+ can also be reabsorbed from the xylem fluid along the pathway to the shoot, mostly in the upper root and lower shoot parts. At stronger salt stress, that is, in an environment of very low water potential, resistant plants must absorb the salt for osmoregulation. This causes problems because large Na^+ levels are not compatible with the cytoplasmic machineries of metabolism. Thus, NaCl is sequestered in the vacuoles of the cells by transport across the tonoplast. To avoid shrinkage and dehydration, and maintain the water structures of the cytoplasm, compatible organic compounds are synthesized and accumulated in the cytoplasm, which often occupies only a few percent of the total cell volume of such rather succulent plants. Compatible solutes are, for example, polyalcohols, quaternary ammonia, and sulfonium compounds. *See* PLANTS OF SALINE ENVIRONMENTS.

Special physiological mechanisms. Growing plants depend not only nutritionally on ion transport. Extension growth of young organs is driven by an

H^+—K^+ exchange mechanism. The proton extrusion pump of cells is stimulated by the phytohormone β-indole acetic acid. K^+ is taken up in exchange for H^+ to provide electrical charge balance. But, without a mechanism regulating pH, the cytoplasm would become unduly alkaline. This is controlled by CO_2 dark fixation giving malic acid via phosphoenolpyruvate and oxaloacetate. The protons of the newly formed carboxyl groups are extruded; the organic anion malate together with K^+ is transported into the vacuoles, acting as osmotically active material building up turgor pressure, which drives cell extension.

In this way, hydrogen ion pumps quite generally can be parts of pH-regulating mechanisms and turgor-dependent processes. Other examples are movements of stomatal guard cells and of pulvini bringing about nyctinastic and seismonastic changes of leaf positions. Turgor changes in the guard cells can operate via H^+—K^+ exchange and malic acid synthesis; increased turgor leads to opening of, and decreased turgor to closing of, stomata.

Movements and other reactions in plants are accompanied by action potentials. The ionic basis for plant action potentials has been investigated in some algae (Characeae, *Acetabularia*). Depolarization and repolarization of the membranes during action potentials are due to efflux of Cl^- followed by K^+. In other cases, Ca^{2+} is also involved. *See* PLANT OF TRANSPORT SOLUTES; PLANT-WATER RELATIONS.

Ulrich Lüttge

Bibliography. F. M. Harold, The 1978 Nobel Prize in chemistry, *Science*, 202(4373), December 15, 1978; N. Higinbotham, Electropotential of plant cells, *Annu. Rev. Plant Physiol.*, 24:25–46, 1973; U. Lüttge and N. Higinbotham, *Transport in Plants*, 1979; U. Lüttge and M. G. Pitman (eds.), Transport in plants II, part A: Cells, part B: Organs and tissues, *Encyclopedia of Plant Physiology*, *New Series*, 1976; M. G. Pitman, Ion transport in the xylem, *Annu. Rev. Plant Physiol.*, 28:71–88, 1977; L. Reinhold and A. Kaplan, Membrane transport of sugars and amino acids, *Annu. Rev. Plant Physiol.*, 35:45–83, 1984.

Ionic crystals

A class of crystals in which the lattice-site occupants are charged ions held together primarily by their electrostatic interaction. Such binding is called ionic binding. Empirically, ionic crystals are distinguished by strong absorption of infrared radiation, good ionic conductivity at high temperatures, and the existence of planes along which the crystals cleave easily. *See* CRYSTAL STRUCTURE.

Compounds of strongly electropositive and strongly electronegative elements form solids which are ionic crystals, for example, the alkali halides, other monovalent metal halides, and the alkaline-earth halides, oxides, and sulfides. Crystals in which some of the ions are complex, such as metal carbonates, metal nitrates, and ammonium salts, may also be classed as ionic crystals.

As a crystal type, ionic crystals are to be distinguished from other types such as molecular crystals, valence crystals, or metals. The ideal ionic crystal as defined is approached most closely by the alkali halides. Other crystals often classed as ionic have binding which is not exclusively ionic but includes a certain admixture of covalent binding. Thus the term ionic crystal refers to an idealization to which real crystals correspond to a greater or lesser degree, and crystals exist having characteristics of more than one crystal type. *See* CHEMICAL BONDING.

Ionic crystals, especially alkali halides, have played a very prominent role in the development of solid-state physics. They are relatively easy to produce as large, quite pure, single crystals suitable for accurate and reproducible experimental investigations. In addition, they are relatively easy to subject to theoretical treatment since they have simple structures and are bound by the well-understood Coulomb force between the ions. This is in contrast to metals and covalent crystals, which are bound by more complicated forces, and to molecular crystals, which either have complicated structures or are difficult to produce as single crystals. Being readily available and among the simplest known solids, they have thus been a frequent and profitable meeting place between theory and experiment. These same features of ionic crystals have made them attractive as host crystals for the study of crystal defects: deliberately introduced impurities, vacancies, interstitials, and color centers. *See* COLOR CENTERS; CRYSTAL DEFECTS.

Crystal structure. The simplest ionic crystal structures are those of the alkali halides. At standard

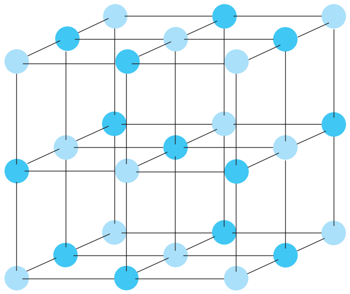


Fig. 1. Sodium chloride lattice. The darker circles represent positive ions and the lighter circles negative ions. (*After F. Seitz, The Modern Theory of Solids, Dover, 1987*)

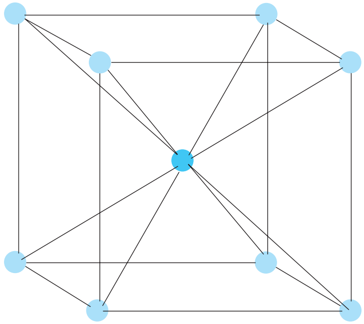


Fig. 2. Cesium chloride lattice. (*After F. Seitz, The Modern Theory of Solids, Dover, 1987*)

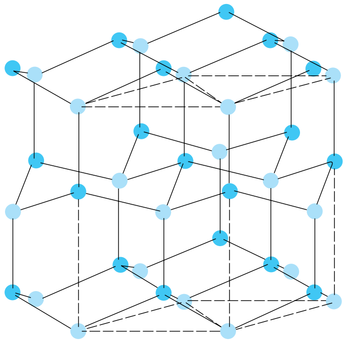


Fig. 3. Wurtzite lattice. (*After F. Seitz, The Modern Theory of Solids, Dover, 1987*)

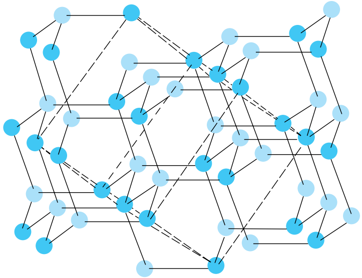


Fig. 4. Zinc blende lattice. (*After F. Seitz, The Modern Theory of Solids, Dover, 1987*)

temperature and pressure the 16 salts of lithium (Li), sodium (Na), potassium (K), and rubidium (Rb) with fluorine (F), chlorine (Cl), bromine (Br), and iodine (I), have the sodium chloride structure of interpenetrating face-centered cubic lattices (**Fig. 1**). Cesium fluoride (CsF) also has this structure but otherwise the cesium halides have the cesium chloride structure of interpenetrating simple cubic lattices (**Fig. 2**). The sodium chloride structure is also assumed by the alkaline-earth oxides, sulfides, and selenides other than those of beryllium (Be) and by silver fluoride (AgF), silver chloride (AgCl), and silver bromide (AgBr). Other crystal structures, such as the wurtzite structure (**Fig. 3**) assumed by beryllium oxide (BeO), β-zinc sulfide (β-ZnS, also known as wurtzite), and zinc oxide (ZnO) and the zinc blende structure (**Fig. 4**) assumed by copper(I) chloride (CuCl), copper(I) bromide (CuBr), copper(I) iodide (CuI), beryllium sulfide (BeS), and α-zinc sulfide (α-ZnS, also known as sphalerite or zinc blende) are also typical of the ionic crystals of salts in which the atoms have equal positive and negative valence. Ionic compounds consisting of monovalent with divalent elements crystallize typically in the fluorite structure (**Fig. 5**) assumed by calcium fluoride (CaF_2), barium fluoride (BaF_2), cadmium fluoride (CdF_2), lithium oxide (Li_2O), lithium sulfide (I_2S), sodium monosulfide (Na_2S), copper(I) sulfide (Cu_2S), and copper(I) selenide or the rutile structure (**Fig. 6**) assumed by titanium(IV) oxide (TiO_2, also known as rutile), zinc fluoride (ZnF), and magnesium fluoride (MgF_2).

Cohesive energy. It is possible to understand many of the properties of ionic crystals on the basis of a simple model originally proposed by M. Born and E. Madelung. In the simplest form of this model, the lattice sites are occupied by spherically symmetric ions having charges corresponding to their normal chemical valence. These ions overlap only slightly the ions at neighboring sites and interact with one another through central forces. In sodium chloride (NaCl), for example, the spherically symmetric closed shell configurations which the free Na^+ and Cl^- ions possess are considered to be negligibly altered by the crystalline environment and to have charges $+e$ and $-e$, respectively, where $-e$ is the charge on the electron. Using this model, together with certain assumptions about the forces between the ions, Born and M. Göppert-Mayer calculated the cohesive energy of a number of ionic crystals. This cohesive energy is defined as the energy necessary to take an ionic crystal from an initial stage, in which the crystal is at 0 K (−459.67°F) and zero pressure. While it cannot be measured directly, this cohesive energy can be deduced from experimental quantities by the use of the Born-Haber cycle. Thus the validity of the simple model of Born and Madelung can be tested by comparing the calculated cohesive energy of Born and Mayer with values which have been experimentally determined. *See* COHESION (PHYSICS).

Born-Haber cycle. This is a sequence of processes leading from the initial to the final state specified in the definition of the cohesive energy. Because in most of the processes in this cycle heat changes at constant pressure are measured, it is convenient to consider the change in heat content or enthalpy

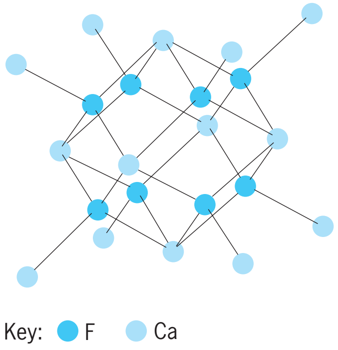


Fig. 5. Calcium fluoride lattice. (*After F. Seitz, The Modern Theory of Solids, Dover, 1987*)

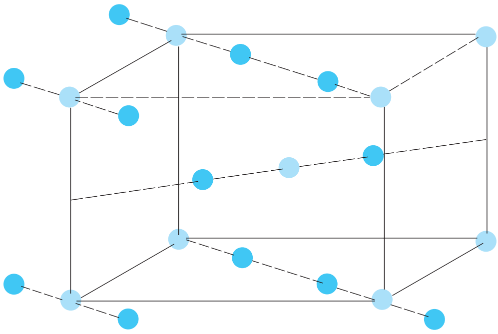


Fig. 6. Rutile lattice. (*After F. Seitz, The Modern Theory of Solids, Dover, 1987*)

$H = U + PV$, where P is the pressure, V is the volume, and U is the cohesive energy, rather than the change in E in each step. Since the total change in H is independent of the intermediate steps taken to accomplish it, the ΔH for the change in state specified in the definition of cohesive energy will be given by the sum of the ΔH in the steps of the cycle. Furthermore, because when $P = 0$, $\Delta H = \Delta U$, the ΔH thus calculated will be the cohesive energy. In the following enumeration of the steps in the Born-Haber cycle $[A]$ indicates element A in a monatomic gaseous state and $[A]$ indicates A in a solid state, and so on. The B without brackets in step 3 refers to the natural form of B at the given temperature and pressure. The ionic compound is taken to be AB, where A is the electropositive and B the electronegative element.

The steps of the Born-Haber cycle (all temperatures in K and pressures in atmospheres) are:

1. $[AB]^{0^\circ}_{P=1} \rightarrow [AB]^{0^\circ}_{P=1}$. The value of ΔH_1 in this isothermal compression is very small and can be neglected in comparison with other heat content changes in the cycle.

2. $[AB]^{0^\circ}_{P=1} \rightarrow [AB]^{298^\circ}_{P=1}$. In this step the crystal is warmed to room temperature. The value of ΔH_2 can be calculated from the specific heat at constant pressure for the crystal.

3. $[AB]^{0^\circ}_{P=1} \rightarrow [A]^{298^\circ}_{P=1} + B^{298^\circ}_{P=1}$. The value of ΔH_3 is given by the heat of formation of the compound AB which is referred to substances in their natural forms at standard temperature and pressure.

4. $B^{298^\circ}_{P=1} \rightarrow (B)^{298^\circ}_{P=1}$. The value of ΔH_4 is the dissociation energy necessary to form a monatomic gas from B in its natural state at standard temperature and pressure. For chlorides, for example, this is the dissociation energy of a Cl_2 molecule into Cl atoms.

5. $[A]^{298^\circ}_{P=1} \rightarrow (A)^{298^\circ}_{P=1}$. In this step, ΔH_5 is the heat of sublimation of the metal A. It can be deduced from the heat of fusion, the specific heats of the solid, liquid, and gaseous phases, and the vapor pressure data for the metal.

6. $(A)^{298^\circ}_{P=1} \rightarrow (A)^{0^\circ}_{P=0}$; $(B)^{298^\circ}_{P=1} \rightarrow (B)^{0^\circ}_{P=0}$. An adiabatic expansion of the gases, considered as ideal, to a very large volume results in a state in which $P = 0$, $T = 0$, and $\Delta H_6 = -{}^{5}\!/\!\sqrt{2}\, RT$/mole, where R is the gas constant.

7. $(A)^{0^\circ}_{P=0} \rightarrow (A^+)^{0^\circ}_{P=0} + e^-$. The ionization of the A atoms gives a ΔH_7 per atom equal to their first ionization energy.

8. $(B)^{0^\circ}_{P=0} + e^- \rightarrow (B^-)^{0^\circ}_{P=0}$. The electrons from step (7) are placed on the B atoms. The value of ΔH_8 per atom is given by the electron affinity of the B atom. *See* ELECTRONEGATIVITY; SUBLIMATION.

As an example, for sodium chloride $\Delta H_1 \cong 10^{-4}$ (kilocalories/mole), $\Delta H_2 = 2.4$, $\Delta H_3 = 98.3$, $\Delta H_4 = 26.0$, $\Delta H_5 = 28.8$, $\Delta H_6 = -2.9$, $\Delta H_7 = 11.9$, and $\Delta H_8 = -80.5$. Experimental cohesive energies for a number of the ionic crystals are given in the **table**. *See* STRUCTURAL CHEMISTRY.

Born-Mayer equation. By use of the Born-Madelung model, the cohesive energy of an ionic crystal can be related to its measured compressibility and lattice spacing. Because of the opposite signs of electric charges which they carry, the unlike ions in such a crystal model attract one another according to Coulomb's law. However, such a charge distribution cannot be in equilibrium if only Coulomb forces act. In addition to their Coulomb interaction, the ions exhibit a repulsion which, compared with the Coulomb interaction, varies rapidly with interionic separation. The repulsion becomes strong for small separations and diminishes rapidly for increasing separation. The static equilibrium configuration of the crystal is determined by a balance of these forces of attraction and repulsion. *See* COULOMB'S LAW.

Cohesive energies*

Crystal	Structure	U_{exp}, kcal/mole†	U_{calc}, kcal/mole†	U_{calc} refined, kcal/mole†
LiCl	NaCl	201.5	196.3	200.2
LiBr	NaCl	191.5	184.4	189.5
LiI	NaCl	180.0	169.1	176.1
NaCl	NaCl	184.7	182.0	183.5
NaBr	NaCl	175.9	172.7	175.5
NaI	NaCl	166.3	159.3	164.3
KCl	NaCl	167.8	165.7	167.9
KBr	NaCl	161.2	158.3	161.3
KI	NaCl	152.8	148.2	152.4
RbCl	NaCl	163.6	159.1	162.0
RbBr	NaCl	159.0	151.9	156.1
RbI	NaCl	149.7	143.1	148.0
CaF_2	Fluorite	618.0	617.7	
CuCl	Zinc blende	226.3	206.1	
ZnS	Wurtzite	851	816	
PbO_2	Rutile	2831	2620	
AgCl	NaCl	207.3	187.3	

*The cohesive energies in the last two columns are calculated using the Born-Mayer equation and the refined Born-Mayer theory, respectively. The refined calculations for the last five crystals have not been made.

†1 kcal = 4.184 kJ.

The short-range repulsion between ions must be described by quantum mechanics. When the electron orbits of two ions overlap, the electron charge density in the region of overlap is diminished as a consequence of the Pauli exclusion principle. This charge redistribution results in a repulsion between the ions in addition to the Coulomb interaction, which they have at all interionic distances. In early work, the energy V_{rep} due to repulsion of two ions at a distance r was assumed to have the form of Eq. (1), where B and n are constants to be determined.

$$V_{rep} = \frac{B}{r^n} \qquad (1)$$

Quantum-mechanical calculations of the interaction of atoms with closed shells of electrons indicate that the interaction of repulsion is better approximated by an exponential dependence on interionic distance given in Eq. (2), where A and ρ are constants.

$$V_{rep} = Ae^{-r/\rho} \qquad (2)$$

Both forms for V_{rep} give almost the same calculated cohesive energy; the exponential form gives slightly better agreement with experiment. *See* EXCLUSION PRINCIPLE.

Using the exponential form for the repulsive interaction energy, the potential energy $\varphi(r_{ij})$ of a pair of ions i and j can be written as Eq. (3), where Z_ie

$$\varphi(r_{ij}) = \frac{Z_iZ_je^2}{r_{ij}} + Ae^{-r_{ij}/\rho} \qquad (3)$$

and Z_je are the net charges of the ions i and j, and r_{ij} is the distance of separation of their centers. The assumption that the ions are spherically symmetric has been used here in writing the Coulomb interaction as that of point charges.

The cohesive energy U of an ionic crystal due to the Coulomb and repulsive interactions of its ions is the sum taken over all pairs of ions in the crystal as in Eq. (4), where in the summation the lattice site

$$U = \frac{1}{2}\sum_{i,j}{}' \varphi(r_{ij}) \qquad (4)$$

indices, i and j, range over all sites of the crystal. The prime on the summation sign indicates the exclusion from the sum of terms for which $i = j$ and the factor of $^1/_2$ avoids counting pairs of ions twice.

For crystals in which there are only two types of ion, the Coulomb or electrostatic part of U, U_e, can be written in a simple form given in Eq. (5), where

$$U_e = \frac{1}{2}\sum_{ij}{}' \frac{Z_iZ_je^2}{r_{ij}} = -\frac{N\alpha_M(Z_+Z_-)e^2}{r} \qquad (5)$$

$+ Z_+e$ and $-Z_-e$ are the charges of the positive and negative ions, N is the number of ion pairs in the crystal, r is the nearest anion-cation separation, and α_M is the Madelung constant. *See* MADELUNG CONSTANT.

By anticipating that ρ will be small compared to the nearest neighbor separation, the interactions of repulsion may be neglected for pairs of ions other than nearest neighbors. The energy of the crystal model for arbitrary nearest neighbor separation r is then given by Eq. (6), where M is the number of

$$U(r) = N\frac{-\alpha_M Z_+Z_-e^2}{r} + MAe^{-r/\rho} \qquad (6)$$

nearest neighbors which each ion has in the crystal.

The parameter ρ may be evaluated for a given crystal by requiring that (1) U be a minimum for the observed value of r and (2) that the compressibility of the model equal the measured compressibility of the crystal. It follows from these requirements that Eq. (7) holds. This is the Born-Mayer equation for

$$U(r_0) = \frac{N\alpha_M Z_+Z_-(1-\rho/r_0)e^2}{r_0} \qquad (7)$$

the cohesive energy, where r_0 refers to the nearest neighbor distance at static equilibrium. Further, in this equation, ρ is given in terms of experimental quantities by Eq. (8), where K is the measured compressibility of the crystal.

$$\frac{r_0}{\rho} = \frac{18r_0^4}{\alpha_M e^2 K} + 2 \qquad (8)$$

Cohesive energies for some alkali halides and crystals of other structures calculated in this way are shown in the table, where they can be compared with the experimental values for the cohesive energy. The agreement is considered to be support for the essential validity of the Born-Madelung model. The model has been applied with some success even to the ammonium halides, assuming spherically symmetric ions.

The Born-Mayer theory has been refined, with resulting improvement in the agreement between the calculated and experimental cohesive energies for alkali halides. The refinements have considered the small (a few kilocalories per mole or less) corrections to the cohesive energy arising from van der Waals interactions and zero-point vibrational energy. The van der Waals forces are weak attractive forces between ions due to mutually induced fluctuating dipoles. Similar forces, even weaker, due to dipole-quadrupole interactions have also been considered. Both these interactions make small positive contributions to the cohesive energy. At 0 K (−459.67°F) the lattice is not in static equilibrium but as a consequence of quantum mechanics is in a state of zero-point vibration with nonzero energy. The energy of these vibrational modes cannot be further reduced. The zero-point vibration energy gives a small negative contribution to the cohesive energy. The results of these refinements of the Born-Mayer theory are also shown in the table. *See* LATTICE VIBRATIONS.

While it has had success in enabling the cohesive energy to be calculated, the shortcomings of this simple model become evident in its failure to predict correctly the elastic shear constants of ionic crystals. This requires interionic forces of a noncentral character, which are absent from the model. There are also other instances in which the simple model is found to be inadequate.

More elaborate models which take into account features absent from the Born-Mayer model and which may be regarded as extensions of it have had considerable success in accounting for the elastic and dielectric properties of alkali halides. More important has been the ability of these models to account for the lattice phonons in these crystals. This has given the alkali halides a prominent place in the study of phonons in insulators, where they are among the few reasonably well-understood solids. B. Gale Dick

Ionic conductivity. Most ionic crystals have large band gaps, and are therefore generally good electronic insulators. However, electrical conduction occurs by the motion of ions through these crystals. The presence of point defects, that is, deviations from ideal order in the crystalline lattice, facilitates this motion, thus giving rise to transport of electric charge. In an otherwise perfect lattice where all lattice sites are fully occupied, ions cannot be mobile.

Mechanism. The two most common types of point defects that give rise to conductivity in ionic crystals are vacancies and interstitials. These intrinsic defects are in thermodynamic equilibrium, and their concentrations depend exponentially upon the temperature. The most common mechanisms by which ions are conducted are vacancy diffusion and interstitial diffusion. In the former, a mobile ion positioned close to a vacancy may move into it. In the latter, the interstitial ion can change sites by hopping. Ionic crystals with Frenkel disorders (cation vacancy-cation interstitial pairs), such as silver halides, where the predominant defects are silver interstitials, have substantially greater values of ionic conductivity than those with Schottky disorders (anion-cation vacancy pairs), such as alkali halides and alkaline-earth halides, where the predominant defects are vacancies.

Vacancies and interstitials can also be generated extrinsically by doping the ionic crystal with a suitable amount of an aliovalent (that is, of different valence from the host ion) species. This leads to the formation of vacancies or interstitial defects to counterbalance the excess charge brought in by the aliovalent ion since the crystal must preserve its charge neutrality.

Superionic conductors. Many so-called normal ionic crystals possess conductivities of about 10^{-10} (ohm·cm)$^{-1}$ or lower at room temperature. However, a relatively small number of ionic materials, called superionic conductors or fast ionic conductors, display conductivities of the order of 10^{-1} to 10^{-2} (ohm · cm)$^{-1}$, which imply ionic liquidlike behavior. In most of these crystals, only one kind of ionic species is mobile, and its diffusion coefficient and mobility attain values such as found otherwise only in liquids. Due to their high value of ionic conductivity as well as their ability to selectively transport ionic species, superionic conductors have successfully been employed as solid electrolytes in many applications. *See* DIFFUSION; SOLID-STATE BATTERY.

The high value of conductivity and diffusion coefficient for the mobile ionic species is due to the high degree of structural disorder in these materials. This disorder is so great that the mobile ions assume a quasimolten state, swimming in the open spaces in a rigid framework of the immobile ions. For example, the high-temperature α-phase of silver iodide (α-AgI), stable above 147°C (297°F), attains unusually high values of ionic conductivity, in excess of 1 (ohm · cm)$^{-1}$. The iodide ions occupy the body-centered cubic (bcc) sites and form the rigid framework. The silver ions are distributed among sites whose numbers are much greater than the number of silver ions. Similarly, $RbAg_4I_5$ has 56 tetrahedral interstitial positions for the 16 silver ions present in the structure to occupy. [This material possesses the highest ionic conductivity yet reported for a solid at room temperature, 0.3 (ohm · cm)$^{-1}$ at 300 K or 80°F.]

The preexponential factor in the Arrhenius expression for the temperature dependence of the conductivity is appreciably smaller in superionic conductors than that exhibited by normal ionic materials. The activation enthalpy for conduction is generally small in superionic conductors, but can range over a wide spectrum. For example, doped zirconium dioxide (ZrO_2), which transports oxide ions by vacancy diffusion and is used at elevated temperatures between 600 and 1400°C (1100 and 2500°F), possesses a high activation enthalpy for ionic conduction. The oxygen vacancies in the lattice are produced by dissolution of dopants such as calcium oxide (CaO), magnesium oxide (MgO), or yttrium oxide (Y_2O_3) in zirconium dioxide. Since the Ca^{2+} ion substitutes for the Zr^{4+} site, one oxygen vacancy is produced for every Ca^{2+} ion. The typical amount of doping in this material is of the order of 10 mol %. This results in a very high concentration of oxygen vacancies of about 10^{22} cm^{-3}, comparable to the electron density in metals. *See* ENTHALPY.

Parameters and measurements. If charge and mass are transported by the same mobile species, then independent measurements of the ionic conductivity and the diffusion coefficient allow calculation of the concentration as well as the electrical mobility of that species. Even the most mobile ions in superionic conductors have mobilities that are orders of magnitude lower than that of electrons in metals. Hence the ionic conductivity of superionic conductors is still much below that of metals.

An important consideration is the extent to which a mobile ionic species contributes to the total conduction process. The transference number t_i for a species i is defined as the ratio of the partial (or specific) conductivity due to species i to the total conductivity. It can be determined by measuring the open-circuit potential across the ionic crystal in an electrochemical cell arrangement, such that a known difference in the chemical potential of species i is imposed on each side of the crystal at a suitable temperature. The ratio of the measured open-circuit potential to the theoretical potential calculated from the known chemical potential difference gives the transference number. Another method is to use an electrochemical titration technique with suitable slab electrodes that contain the mobile species. With passage of current, the mobile species i is tranported across the ionic material. The weight change in the electrode material due to deposition or depletion of i is compared with the expected value calculated from Faraday's law. However, in all these measurements the interface between the electrode and the ionic material is very important. *See* CHEMICAL THERMODYNAMICS; ELECTROCHEMICAL TECHNIQUES; ELECTROLYSIS.

In order to measure the ionic conductivity or the transference number of the mobile species i in an ionic material, an electrode material that contains species i is required to obtain an interface with the ionic material that is completely nonblocking (sometimes called reversible) to species i but completely blocking (that is, nonreversible) to all other species. By using suitable electrode materials the conductivities and the transference numbers of the minority carriers as well as the mobile species can be obtained.

The easiest way to measure ionic conductivity is a two-probe dc technique with an electrochemical cell arrangement that has nonblocking interfaces for the species i on each side of the ionic material. In practice, there may be problems with slow electrode reactions that must be isolated and accounted for. A better way is the four-probe dc technique using four nonblocking electrodes, two as voltage probes and the other pair to pass a known and constant current.

If a suitable nonblocking electrode material is not available, ac techniques may be used. A small-amplitude sinusoidal signal at reasonably high frequencies can be applied to the ionic material in an electrochemical cell arrangement, and the conductivity value extracted from the real part of the measured cell impedance. However, if the frequency used in the measurements is not high enough, the conductivity value may be in error. A more involved technique to eliminate this problem is ac impedance spectroscopy. Here a wide range of frequencies covering many decades is imposed upon the ionic material, and the frequency dispersion is analyzed in the complex impedance plane. Experience and insight are required for meaningful analyses of the frequency spectrum. *See* ALTERNATING-CURRENT CIRCUIT THEORY. Turgut M. Gür

Dielectric properties. The polarization of an ionic crystal in an applied time-varying electric field depends on the frequency v of the field. If $v = 0$ the field is static, and a static equilibrium will be achieved in which the ions are displaced from their lattice sites a distance determined by a balance between the force of repulsion due to their nearest neighbors and the electrical force on the ions. Oppositely charged ions move in opposite directions with a resulting ionic polarization. In addition, the electron clouds of the ions are deformed in the local electric field at the ions and, to a much lesser extent, by the repulsive forces, to give an electronic polarization. As the frequency v increases, inertial effects must be considered. When v is sufficiently high, that is, when $v > v_0$ (v_0 is the natural frequency of vibration of ions about lattice sites), the heavy ions can no longer follow the rapid variations of the applied field. The ionic polarization ceases to contribute to the total polarization, although the less massive electron clouds continue to give an electronic polarization. Eventually, with increasing frequency, v exceeds the frequency at which the electron cloud deformations can follow the field, and the electronic polarization becomes altered. The frequency v_0, called the restrahlung or residual-ray frequency, lies in the infrared region, and the associated vibrations of the charged ions are responsible for the strong absorption of infrared radiation in ionic crystals. The term residual-ray frequency comes from the fact that the reflectivity of visible and infrared radiation is highest for radiation of a frequency near to that of maximum absorption, v_0. Thus multiple reflection will yield radiation composed predominantly of residual rays of frequency near to v_0. For additional information on residual rays *see* REFLECTION OF ELECTROMAGNETIC RADIATION

At frequencies below v_0 and above it up to the ultraviolet, an ideal alkali halide is transparent to electromagnetic radiation. Impurities or defects often introduce absorption bands into these extensive and otherwise transparent regions and can thus be studied optically. This fact accounts for much of the wealth of experimental detail available on defects and impurities in alkali halides and for the interest which they have attracted.

Ionic radii. It has been found that ionic radii can be chosen such that the lattice spacings of ionic crystals are given approximately by the sum of the radii of the constituent ions. Thus, for instance, eight radii suffice to give the approximate lattice spacings of the 16 alkali halides having sodium chloride structure. This is true because of the rapid variation of the repulsive interaction with distance, so that the ions interact somewhat like hard spheres. This implies that a given ion, having the same ionic radius in a variety of crystals, has an electron cloud that is only slightly altered by the differing crystalline environments of these crystals. B. Gale Dick

Bibliography. N. W. Ashcroft and W. D. Mermin, *Solid State Physics*, 1976; H. Boettger et al., *Hopping Conduction in Solids,* 1986; S. Chandra, *Superionic Solids: Principles and Applications,* 1981; C. Kittel, *Introduction to Solid State Physics*, 7th ed., 1996; L. C. Pauling, *The Nature of the Chemical Bond*, 3d ed., 1960; F. Seitz et al., *Solid State Physics*, 40 vols. and 15 suppls., 1955–1987; A. Stoneham (ed.), *Ionic Solids at High Temperatures*, 1989; T. Takahashi (ed.), *High Conductivity in Solid Ionic Conductors: Recent Trends and Applications*, 1989.

Ionic equilibrium

An equilibrium in a chemical reaction in which at least one ionic species is produced, consumed, or changed from one medium to another.

Types of equilibrium. A few examples can illustrate the wide variety of types of ionic equilibrium which are known.

Dissolution of an un-ionized substance. The dissolution of hydrogen chloride (a gas) in water (an ionizing solvent) can be used to illustrate this type. Reactions (1), (2), and (3) all represent exactly the same equilibrium.

$$HCl(g) \rightleftharpoons H^+ + Cl^- \tag{1}$$

$$HCl(g) + H_2O \rightleftharpoons H_3O^+ + Cl^- \tag{2}$$

$$HCl(g) + 4H_2O \rightleftharpoons H_9O_4{}^+ + Cl^- \tag{3}$$

Equation (1) ignores the hydration of the proton and is preferred for many purposes when the hydration (or solvation) of the proton is irrelevant to a particular discussion. Equation (2) is written in recognition of the widely held belief that free protons do not exist in aqueous solution. Equation (3) indicates that another three molecules of water are very firmly bound to the H_3O^+ ion (the hydronium ion). There is no implication in Eq. (3), however, that the total number of molecules of water attached to, or weakly affected by, the hydronium ion may not be considerably larger than three.

Not much is known about the solvation of ions, although it has been proved that each chromic ion, Cr^{3+}, in dilute aqueous solution holds at least six water molecules. Only a few other similar data have been clearly established. Hence equations for the other examples cited below are written without regard to solvation, except that the hydrogen ion is usually written in accordance with common practice as H_3O^+. *See* ACID AND BASE.

Dissolution of a crystal in water. The dissociation of solid silver chloride, reaction (4), illustrates this type of

$$\mathrm{AgCl(crystal)} \rightleftharpoons \mathrm{Ag^+ + Cl^-} \qquad (4)$$

equilibrium. *See* SOLUBILITY PRODUCT CONSTANT.

Dissociation of a strong acid. Nitric acid, HNO_3, dissociates as it dissolves in water, as in reaction (5). At

$$\mathrm{HNO_3 + H_2O \rightleftharpoons H_3O^+ + NO_3{}^-} \qquad (5)$$

77°F (25°C) about one-half the acid is dissociated in a solution containing 10 (stoichiometric) moles of nitric acid per liter.

Dissociation of an ion in water. The bisulfate ion, HSO_4^-, dissociates in water, as in reaction (6). About one-

$$\mathrm{HSO_4{}^- + H_2O \rightleftharpoons H_3O^+ + SO_4{}^{2-}} \qquad (6)$$

half the HSO_4^- is dissociated in an aqueous solution containing about 0.011 mole of sulfuric acid per liter at 77°F (25°C).

Dissociation of water itself. In pure water at 77°F (25°C) the concentration of each ion is about 10^{-7} mole/liter, but increases rapidly as temperature is increased. This equilibrium is represented by reaction (7).

$$\mathrm{2H_2O \rightleftharpoons H_3O^+ + OH^-} \qquad (7)$$

Formation of a complex ion. In water or in a mixture of fused (chloride) salts, complex ions, such as $ZnCl_4{}^{2-}$, may be formed, as in reaction (8).

$$\mathrm{Zn^{2+} + 4Cl^- \rightleftharpoons ZnCl_4{}^{2-}} \qquad (8)$$

See COORDINATION COMPLEXES.

Dissociation of a weak acid. In water acetic acid dissociates to form hydrogen (hydronium) ion and acetate ion, as in reaction (9).

$$\mathrm{CH_3CO_2H + H_2O \rightleftharpoons H_3O^+ + CH_3CO_2{}^-} \qquad (9)$$

Electrochemical reaction. Reaction (10) takes place "al-

$$\mathrm{{}^1\!/_2H_2(\mathit{g}) + AgCl(s) + H_2O(\mathit{l}) \rightleftharpoons H_3O^+ + Cl^- + Ag(s)} \qquad (10)$$

most reversibly" when the equilibrium shown exists. A small current is allowed to flow through an electric cell consisting of an aqueous solution of HCl saturated with silver chloride, a hydrogen electrode, and a silver electrode. Saturation is maintained by an excess of solid silver chloride which for convenience is sometimes mixed with the silver or plated, as a coating, on the metal. The electrode is then called a silver–silver chloride electrode.

Many additional types of equilibria could be mentioned, including those reactions occurring entirely in the gaseous phase and those reactions occurring between substances dissolved in two immiscible liquids.

Quantitative relationships. Each reaction obeys an equilibrium equation of the type shown as Eq. (11).

$$\frac{[\mathrm{H^+}][\mathrm{Cl^-}]}{[\mathrm{HCl}(g)]}\frac{f_+f_-}{\gamma_g} = Q_cQ_f = K \qquad (11)$$

The activity coefficient γ_g can be ignored here because it is very nearly unity. The terms f_+ and f_- are the respective activity coefficients of H^+ and Cl^- but cannot be determined separately. Their product can be determined experimentally and can also be calculated theoretically for very dilute solutions by means of the Debye-Hückel theory of interionic attraction. Because γ_g and f_+f_- are nearly unity, Eq. (11) demands that the pressure of HCl gas above a dilute aqueous solution be proportional to the square of the concentration of the solute. Q_c is called the concentration quotient and Q_f the quotient of activity coefficients.

Similarly, the dissociation of acetic acid obeys Eq. (12), where f_u is the activity coefficient of the

$$\frac{[\mathrm{H^+}][\mathrm{CH_3CO_2^-}]}{[\mathrm{CH_3CO_2H}]}\frac{f_+f_-}{f_u} = Q_cQ_f = K \qquad (12)$$

unionized acetic acid. Early work on electrolytes revealed that Q_c, the concentration quotient, was constant within the limits of accuracy attainable at the time. Later work revealed that the measured concentration quotient Q_c first increases as concentration is increased from very small values and then decreases sharply. The initial increase is due largely to the electrical forces between the ions which reduce the product f_+f_-. There is some evidence that the subsequent decrease in Q_c and the concomitant rise in Q_f are due to removal of some of the monomeric acetic acid by the formation of dimeric acetic acid, $(CH_3CO_2H)_2$. The fact is, however, that knowledge concerning activity coefficients in solutions other than very dilute ones is not yet understood. Even the experimental methods for the measurement of the molecular species involved in some equilibria were not evolved until recently. *See* ELECTROLYTIC CONDUCTANCE; HYDROLYSIS. Thomas F. Young

Bibliography. G. M. Barrow, *Physical Chemistry*, 6th ed., 1996; H. Russotti, *The Study of Ionic Equilibria: An Introduction*, 1978; G. H. Schenk and D. D. Ebbing, *Qualitative Analysis and Ionic Equilibrium*, 2d ed., 1990.

Ionic liquids

In the simplest sense, the liquid phase that forms on the melting of an ionic compound (that is, a salt) and consists entirely of discrete ions (**Fig. 1**). More specifically, ionic liquids are the liquid phase of organic salts (organic, complex cations) which

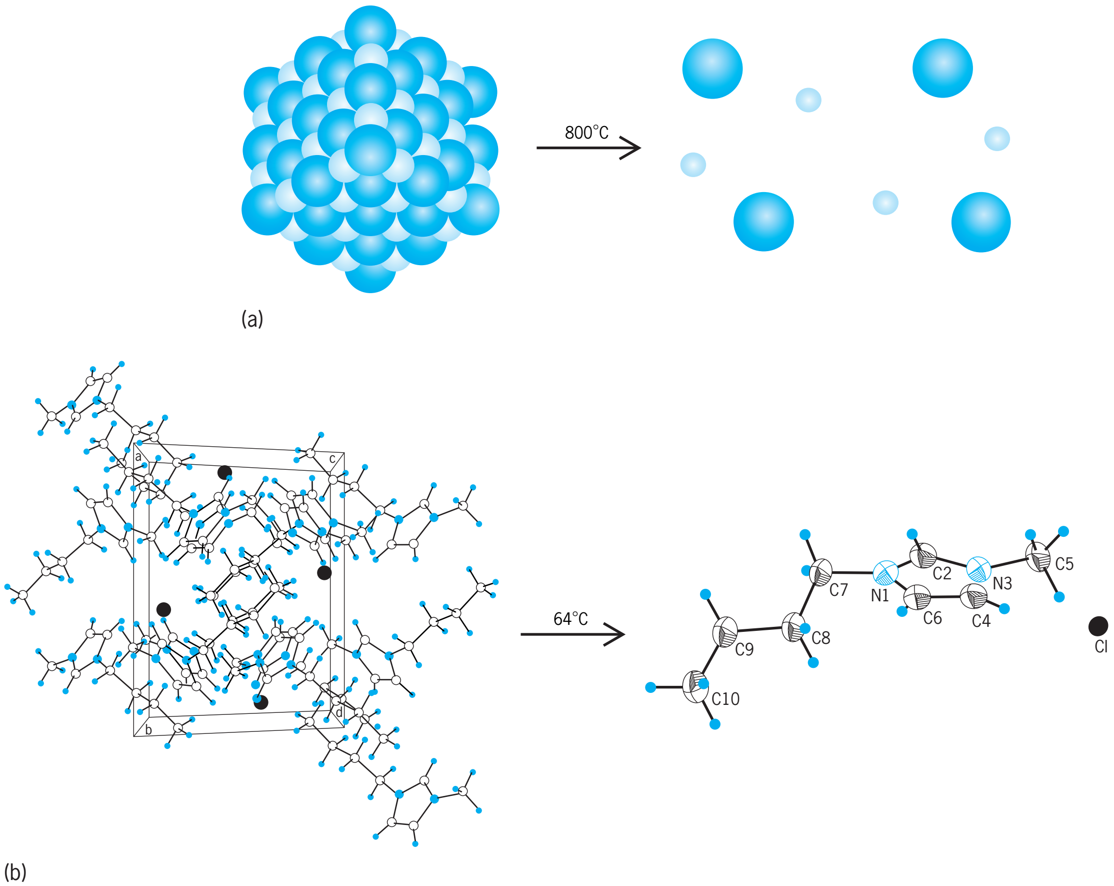
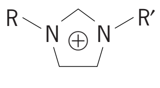


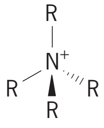
Fig. 1. Examples of (*a*) inorganic and high-melting salts and (*b*) organic and low-melting salts known as ionic liquids. (*After J. D. Holbrey et al., Crystal polymorphism in 1-butyl-3 methylimidazolium halides: Supporting ionic liquid formation by inhibition of crystallization, Chem. Commun., 2003:1636–1637, 2003*)

melt at, or near, ambient temperature (below 100–150°C; 212–302°F). Ionic liquids chemically resemble both phase-transfer catalysts and surfactants and, in some cases, have been used for both purposes. *See* PHASE-TRANSFER CATALYSIS; SALT (CHEMISTRY); SURFACTANT.

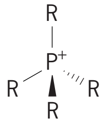
The most extensively studied ionic liquids are systems derived from 1,3-dialkylimidazolium (structure **1**), tetraalkylammonium **(2)**, tetraalkylphosphonium **(3)**, and *N*-alkylpyridinium **(4)** cations.



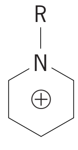
(1)



(2)



(3)



(4)

Derivatives made from these cations can increase or decrease functionality, such as alkyl chains, branching, or chirality. Common anions that allow for the formation of low-melting ionic liquids typically have diffuse (delocalized) charge and can range from simple inorganic anions such as chloride (Cl^-), bromide (Br^-), and iodide (I^-), through larger pseudospherical polyatomic anions, including hexafluorophosphate $[PF_6]^-$ and tetrafluoroborate $[BF_4]^-$, to larger, flexible fluorinated anions such as bis(trifluoromethylsulfonamide) $[(CF_3SO_2)N]^-$ and tri(perfluoroX) trifluorophosphate. A wide range of simple polyhalometallate and halometallate/halide complex anion ionic liquids have also been extensively investigated and have important application in electrochemistry and electrodisposition. The anionic component of the ionic liquid typically controls the solvent's reactivity with water, coordinating ability, and hydrophobicity. Anions can also contain chiral components or can be catalytically active, such as carboranes, polytungstates, and tetrachloroaluminate anions. Manipulation of the rheological properties through mixing of cation-anion pairs to obtain materials that support or enhance reactions makes the use of ionic liquids in organic synthesis an intriguing possibility.

Synthesis. Preparation of the cationic portion of an ionic liquid can be achieved through the quaternization of phosphines or amines with a haloalkane or through protonation with a free acid. Quaternization is generally regarded as the

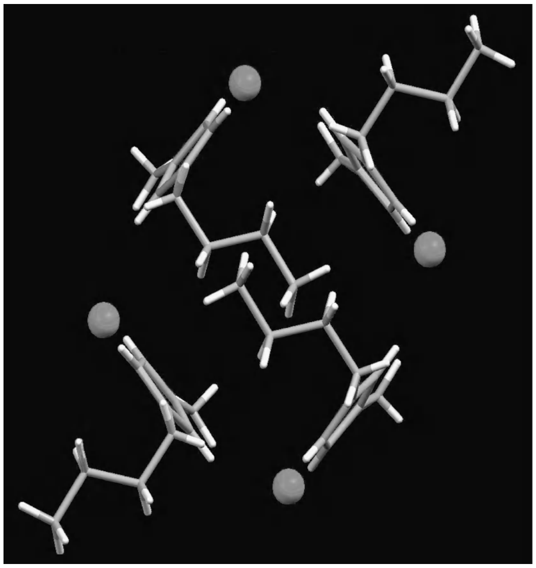


Fig. 2. X-ray crystal structure of 1-butyl-3-methylimidazolium chloride ([C_{4mim}]Cl) in its ortho polymorph (its most common conformation). (*After J. D. Holbrey et al., Crystal polymorphism in 1-butyl-3 methylimidazolium halides: Supporting ionic liquid formation by inhibition of crystallization, Chem. Commun., 2003:1636–1637, 2003. Reproduced by permission of The Royal Society of Chemistry.*)

more sound approach because cations prepared through protonation reactions can be degraded easily through deprotonation, leading to the breakdown of the solvent. Reaction time and temperature for typical quaternization reactions depend upon both the haloalkane and the cation "backbone" used. The most widely used ionic liquids in research include 1-butyl-3-methylimidazolium chloride [C_4mim]Cl, [C_4mim][PF_6], and [C_4mim][BF_4] (**Fig. 2**). Ionic liquids prepared using this method can be modified through anion exchange reactions, acid treatment, or metathesis reactions to prepare ionic liquids containing a desired anion. (The choice of anion is important, as it is the most influential factor on the physical properties of an ionic liquid.) Through these basic reactions, it is possible to prepare a wide range of ionic liquids possessing varied physical properties that can be used to aid in creating optimal reaction conditions. *See* QUATERNARY AMMONIUM SALTS.

Purification of ionic liquids is of utmost importance, specifically those used in organic synthesis, as impurities from preparation could adversely affect the reactions. To synthesize pure ionic liquids, it is necessary to begin with purified starting materials. However, the use of pure starting materials does not ensure pure product. Common impurities in ionic liquids include halide anions from metathesis reactions, the presence of color and water, and acid impurities. Dichloromethane extraction is the most widely accepted method for removing excess halide anions from ionic liquids. Since ionic liquids typically have no measurable vapor pressure, dichloromethane can be removed easily under vacuum. As with organic solvents, water can also be removed from an ionic liquid under vacuum with heat to drive off unwanted moisture. Ionic liquids normally exist as colorless solvents. Impurities that can cause an ionic liquid to become colored can be avoided by using pure starting materials, avoiding the use of acetone in the cleaning of glassware, and keeping temperatures as low as possible during the quaternization step of its synthesis.

Properties. Interest in ionic liquids can be attributed to a number a unique qualities inherent to these materials. First, ionic liquids are excellent conducting materials that can be used in a number of electrochemical applications, including battery production and metal deposition. In fact, they were first designed to be used in electrochemical applications. Second, ionic liquids typically have no measurable vapor pressure, making them an attractive replacement for volatile organic solvents in synthesis. Third, ionic liquids are composed of ions that can be varied to create materials with vastly different physical and chemical properties, compared with conventional solvents. The ability to "fine-tune" the solvent properties of ionic liquids has been exploited to create solvents ideal for a range of organic syntheses. Lastly, ionic liquids exhibit varying degrees of solvation and solubility in a range of organic solvents, allowing simple separations and extractions. *See* SOLVENT.

Use in organic reactions. The replacement of volatile organic solvents with ionic liquids in organic reactions is very exciting. Ionic liquids are polar solvents which can be used as both solvents and reagents in organic synthesis, and the possibility exists that ionic liquids may positively affect the outcome of the reactions.

Certain ionic liquids are suitable for particular classes of organic reactions. For example, neutral ionic liquids are commonly used in Diels-Alder and condensation reactions, as well as nucleophilic displacement. Ionic liquids that possess Lewis acid properties are used in acid-catalyzed reactions, including Friedel-Crafts alkylations and acylations and electrophilic substitutions or additions. *See* ACID AND BASE; ACYLATION; DIELS-ALDER REACTION; FRIEDEL-CRAFTS REACTION; SUBSTITUTION REACTION.

Product isolation from ionic liquids is generally accomplished through extraction with organic solvents. Because most ionic liquids have no measurable vapor pressure, simple vacuum techniques can be used to recover the product in the organic solvent. Biphasic systems composed of ionic liquids and water or organic phases have been used in catalysis reactions exploiting the properties of some ionic liquids to easily recover both product and catalyst. As stated previously, the ionic constituents of an ionic liquid can be "tuned" to exhibit desired physical properties to support such biphasic reactions.

The unique properties of ionic liquids, such as their stability and nonvolatility, make them good candidates for use as solvents in homogeneous catalysis systems. Polymerization reactions in ionic liquids, using transition-metal catalysts and conventional organic initiators, have been studied and have demonstrated increased efficiency compared with traditional polymerization solvents. *See* HOMOGENEOUS CATALYSIS.

Green chemistry. Sustained interest in ionic liquids can be attributed to their desirable physical properties, such as their electron conductivity, as well as their novelty and the possibility of enhanced reactions. The perceived potential to eliminate volatile organic solvents in synthetic and separation processes has also driven interest and investigation into ionic liquids.

Ionic liquids have many characteristics relevant to a general "green chemistry" approach, including the lack of volatility; however, almost all other properties (such as toxicity, stability, and reactivity) vary with the cation and anion components and cannot readily be generalized. The utility and interest in ionic liquids, as a class of fluids, rests with individual examples displaying new, improved, or different combinations of solvent properties. *See* GREEN CHEMISTRY.

M. B. Turner; J. D. Holbrey; S. K. Spear; R. D. Rogers

Bibliography. R. D. Rogers and K. R. Seddon (eds.), *Ionic Liquids; Industrial Applications to Green Chemistry*, ACS Symp. Ser. 818, American Chemical Society, Washington, DC, 2002; R. D. Rogers and K. R. Seddon (eds.), *Ionic Liquids as Green Solvent, Progress and Prospects*, ACS Symp. Ser. 856, American Chemical Society, Washington, DC, 2003; P. Wasserscheid and T. Welton (eds.), *Ionic Liquids in Synthesis*, Wiley-VCH, Weinheim, 2002; T. Welton, Room temperature ionic liquids. Solvents for synthesis and catalysis, *Chem. Rev.*, 99:2071, 1999.

Ionization

The process by which an electron is removed from an atom, molecule, or ion. This process is of basic importance to electrical conduction in gases, liquids, and solids. In the simplest case, ionization may be thought of as a transition between an initial state consisting of a neutral atom and a final state consisting of a positive ion and a free electron. In more complicated cases, a molecule may be converted to a heavy positive ion and a heavy negative ion which are separated.

Ionization may be accomplished by various means. For example, a free electron may collide with a bound atomic electron. If sufficient energy can be exchanged, the atomic electron may be liberated and both electrons separated from the residual positive ion. The incident particle could as well be a positive ion. In this case the reaction may be considerably more complicated, but may again result in a free electron. Another case of considerable importance is the photoelectric effect. Here a photon interacts with a bound electron. If the photon has sufficient energy, the electron may be removed from the atom. The photon is annihilated in the process. Other methods of ionization include thermal processes, chemical reactions, collisions of the second kind, and collisions with neutral molecules or atoms. *See* ELECTRICAL CONDUCTION IN GASES; ELECTRODE POTENTIAL.

Glenn H. Miller

Bibliography. R. Gomer, *Field Emission and Field Ionization*, 1961, reprint 1993; H. Russotti, *The Study of Ionic Equilibria: An Introduction*, 1978; D. Veza (ed.), *Physics of Ionized Gases*, 1995; A. Von Engel, *Ionized Gases*, 2d ed., 1965, reprint 1994.

Ionization chamber

An instrument for detecting ionizing radiation by measuring the amount of charge liberated by the interaction of ionizing radiation with suitable gases, liquids, or solids. These radiation detectors have played an important part in the development of modern physics and have found many applications in basic scientific research, in industry, and in medicine.

Principle of operation. While the gold leaf electroscope (**Fig. 1**) is the oldest form of ionization chamber, instruments of this type are still widely used as monitors of radiation by workers in the nuclear or radiomedical professions. In this device, two thin flexible pieces of gold leaf are suspended in a gas-filled chamber. When these are electrically charged, as in Fig. 1, the electrostatic repulsion causes the two leaves to spread apart. If ionizing radiation is incident in the gas, however, electrons are liberated from the gas atoms. These electrons then drift toward the positive charge on the gold leaf, neutralizing some of this charge. As the charge on the gold leaves decreases, the electrostatic repulsion decreases, and hence the separation between the leaves decreases. By measuring this change in separation, a measure is obtained of the amount of radiation incident on the gas volume. While this integrated measurement may be convenient for applications such as monitoring the total radiation exposure of humans, for many purposes it is useful to measure the ionization pulse produced by a single ionizing particle. *See* ELECTROSCOPE.

The simplest form of a pulse ionization chamber consists of two conducting electrodes in a container

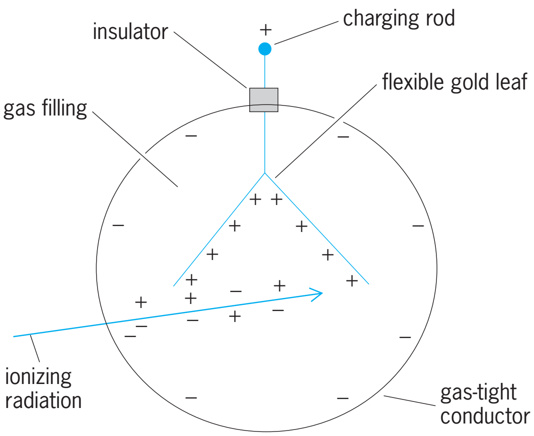


Fig. 1. Gold leaf electroscope used as a radiation detector.

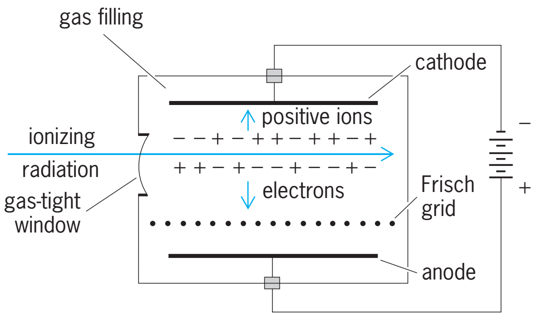


Fig. 2. Parallel-plate ionization chamber.

filled with gas (**Fig. 2**). A battery, or other power supply, maintains an electric field between the positive anode and the negative cathode. When ionizing radiation penetrates the gas in the chamber—entering, for example, through a thin gas-tight window—this radiation liberates electrons from the gas atoms leaving positively charged ions. The electric field present in the gas sweeps these electrons and ions out of the gas, the electrons going to the anode and the positive ions to the cathode.

The basic ion chamber signal consists of the current pulse observed to flow as a result of this ionization process. Because the formation of each electron-ion pair requires approximately 30 eV of energy on the average, this signal is proportional to the total energy deposited in the gas by the ionizing radiation.

Because the charge liberated by a single particle penetrating the chamber is small, very low-noise high-gain amplifiers are needed to measure this charge. In the early days, this was a severe problem, but such amplifiers have become readily available with the development of modern solid-state electronics. *See* AMPLIFIER.

In a chamber, such as that represented in Fig. 2, the current begins to flow as soon as the electrons and ions begin to separate under the influence of the applied electric field. The time it takes for the full current pulse to be observed depends on the drift velocity of the electrons and ions in the gas. These drift velocities are complicated functions of gas type, voltage, and chamber geometry. However, because the ions are thousands of times more massive than the electrons, the electrons always travel several orders of magnitude faster than the ions. As a result, virtually all pulse ionization chambers make use of only the relatively fast electron signal. The electron drift velocities for a few gases are given in **Fig. 3**. Using one of the most common ion chamber gases—argon with a small amount of methane—with electrode spaces of a few centimeters and voltages of a few hundred volts, the electron drift time is of order a microsecond, while the positive-ion drift time is of order milliseconds. By using narrow-bandpass amplifiers sensitive only to signals with rise times of order a microsecond, only the electron signals are observed.

Energy spectrum. One of the most important uses of an ionization chamber is to measure the total energy of a particle or, if the particle does not stop in the ionization chamber, the energy lost by the particle in the chamber. When such an energy-sensitive application is needed, a simple chamber geometry such as that shown in Fig. 2 is not suitable because the fast electron signal charge is a function of the relative distance that the ionization event occurred from the anode and cathode. If an ionization event occurs very near the cathode, the electrons drift across the full electric potential V_0 between the chamber electrodes, and a full electron current pulse is recorded; if an ionization event occurs very near the anode, the electrons drift across a very small electric potential, and a small electron pulse is recorded. This geometrical sensitivity is a result of image charges induced by the very slowly moving positive ions. It can be shown that if the electrons drift through a potential difference ΔV, the fast electron charge pulse is $q' = (\Delta V/V_0)q$, where q is the total ionization charge liberated in the gas.

This geometrical dependence can be eliminated by introducing a Frisch grid as indicated in **Fig. 4**. This grid shields the anode from the positive ions and, hence, removes the effects of the image charges. By biasing the anode positively, relative to the grid, the electrons are pulled through the grid and collected on the anode. Now no signal is observed on the anode until the electrons drift through the grid, but the signal charge which is then observed is the full ionization charge q.

While the ionization chamber generates only small quantities of signal charge for incident particles or photons of megaelectronvolt energies, the resulting signals are nevertheless well above the noise level of modern low-noise electronic amplifiers. When the signals generated by many incident particles of the same energy are individually measured, and a histogram is plotted representing the magnitude of a signal pulse versus the total number of pulses with that magnitude, then an energy spectrum results.

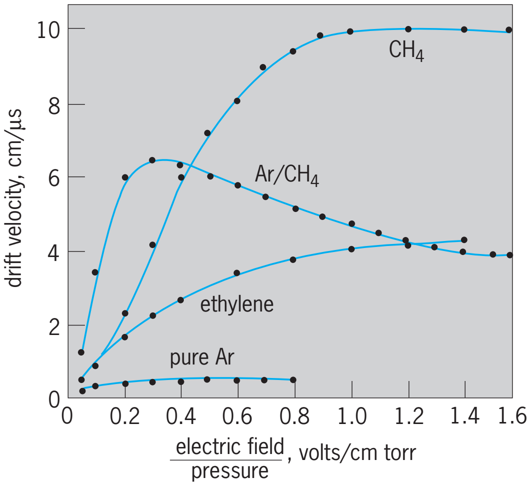


Fig. 3. Electron drift velocity in four different gases as a function of the ratio of the applied electric field strength in volts/centimeter to gas pressure in torrs. 1 torr = 133 Pa. (*After H. W. Fulbright, Ionization chambers, Nucl. Instrum. Meth., 162(1979):21–28,1979*)

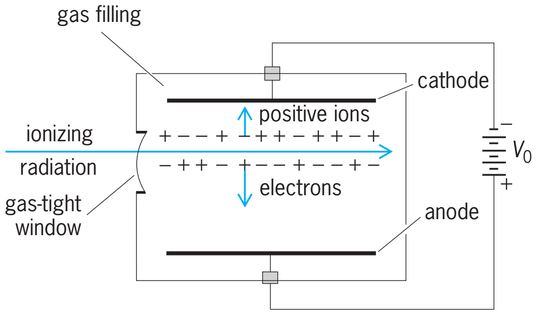


Fig. 4. Frisch grid parallel-plate ionization chamber.

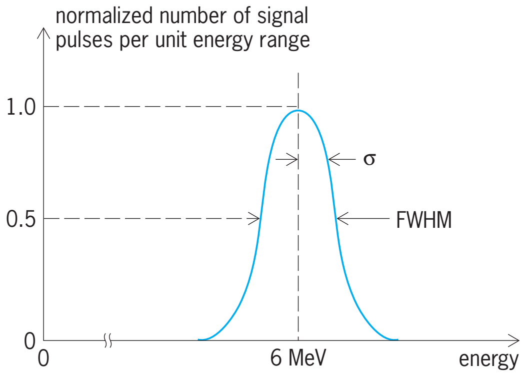


Fig. 5. Idealized energy spectrum produced by monoenergetic (6-MeV) alpha particles incident on an ideal gridded ionization chamber. The spectrum consists of a single gaussian peak with standard deviation σ.

Such a spectrum, smoothed out, consists of an essentially gaussian distribution with standard deviation σ (**Fig. 5**). Assuming a negligible contribution from amplifier noise, it might at first sight appear that σ should correspond to the square root of the average number of electron-ion pairs produced per incident particle. In fact, σ is usually found to be less than this by a substantial amount, usually designated F, where F is the Fano factor. *See* PROBABILITY; STATISTICS.

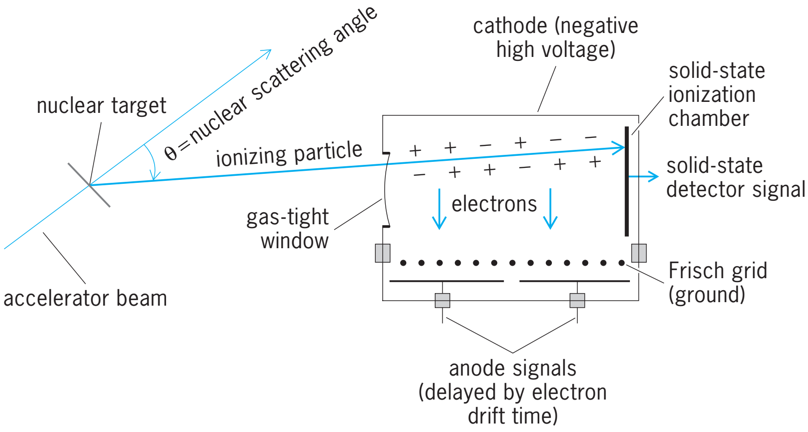


Fig. 6. Heavy-ion detector telescope used to study nuclear reactions.

It is usual to express the width of an energy distribution such as that of Fig. 5 not in terms of σ but in terms of the “full width at half maximum,” usually designated FWHM, or Δ. It can be shown that, for situations in which the width of the energy spectrum is governed by statistics alone, the FWHM is given by the equation $\sigma = 2.36F\epsilon E$, where ϵ is the average energy required to create an electron-ion pair and E is the energy deposited in the chamber by each incident particle or photon. Values of the Fano factor F as low as 0.1 have been observed for certain gases.

Gaseous ionization chambers. Because of the very few basic requirements needed to make an ionization chamber (namely, an appropriate gas with an electric field), a wide variety of different ionization chamber designs are possible in order to suit special applications. In addition to energy information, ionization chambers are now routinely built to give information about the position within the gas volume where the initial ionization event occurred. This information can be important not only in experiments in nuclear and high-energy physics where these position-sensitive detectors were first developed, but also in medical and industrial applications.

This position-sensitivity capability results from the fact that, to a good approximation, electrons liberated in an ionizing event drift along the electric field line connecting the anode and cathode, and they drift with uniform velocity (Fig. 3). Hence, a measure of the drift time is a measure of the distance from the anode that the ionization occurred. A simple illustration of this use is a heavy-ion nuclear physics detector “telescope” used in the basic study of nuclear reactions (**Fig. 6**). Ionizing charged particles (such as ^{1}H, ^{4}He, and ^{12}C) produced in nuclear collisions enter the detector telescope through a thin gas-tight window at the left, pass through two Frisch grid ionization chambers, and then stop in a solid-state ionization detector. Measurement of the ionization produced in the gas versus the total energy of the particle as determined by the solid-state ionization detector gives sufficient information to uniquely identify the mass and atomic charge of the incident particle. Because the response of the solid-state detector is fast relative to the electron drift time, the difference in time of the signals from the solid-state detector and the anode determines the electron drift time and, hence, the distance above the grid that the particle entered the ionization chamber. This distance can be used to determine the nuclear scattering angle. Hence, a very simple device can be designed which gives several pieces of useful information. While this example illustrates the principles, very complex ionization chambers are now routinely used in heavy-ion and high-energy physics where tens or a hundred signals are recorded (using a computer) for a single ionization event. Position-sensitive heavy-ion detectors with active surfaces as large as a square meter have been developed.

Aside from applications in basic nuclear physics research, ionization chambers are widely used in other applications. Foremost among these is the use of gas ionization chambers for radiation monitoring. Portable instruments of this type usually employ a detector containing approximately 1 liter of gas, and operate by integrating the current produced by the ambient radiation. They are calibrated to read out in convenient units such as milliroentgens per hour.

Another application of ionization chambers is the use of air-filled chambers as domestic fire alarms. These employ a small ionization chamber containing a low-level radioactive source, such as ^{241}Am, which generates ionization at a constant rate, the resulting current being monitored by a small solid-state electronic amplifier. On the introduction of smoke into the chamber (which is open to the ambient air), the drifting charged ions tend to attach themselves to the smoke particles. This reduces the ionization chamber current, since the moving charge carriers are now much more massive than the initial ions and therefore exhibit correspondingly reduced mobilities. The observed reduction in ion current is used to trigger the alarm. *See* FIRE DETECTOR.

Another development in ion chamber usage was that of two-dimensional imaging in x-ray medical applications to replace the use of photographic plates. This imaging depends on the fact that if a large flat parallel-plate gas ionization chamber is illuminated with x-rays (perpendicular to its plane), the resulting charges will drift to the plates and thereby form an "image" in electrical charge of the point-by-point intensity of the incident x-rays. This image can be recorded xerographically by arranging for one plate to be a suitably charged insulator. This insulator is preferentially discharged by the collected ions. The resulting charge pattern is recorded by dusting the insulator with a fine powder and transferring this image to paper in the usual xerographic technique. Alternatively, the xerographic insulator may be a photoconductor, such as selenium, which is preferentially discharged by the ionization produced in the solid material. This is then an example of a solid ionization chamber, and its action closely parallels the operation of the optical xerographic copying machines. Such x-ray imaging detectors provide exceedingly high-quality images at a dosage to the patient substantially less than when photographic plates are used. *See* PHOTOCOPYING PROCESSES.

Gaseous ionization chambers have also found application as total-energy monitors for high-energy accelerators. Such applications involve the use of a very large number of interleaved thin parallel metal plates immersed in a gas inside a large container. An incident pulse of radiation, due for example to the beam from a large accelerator, will produce a shower of radiation and ionization inside the detector. If the detector is large enough, essentially all of the incident energy will be dissipated inside the detector (mostly in the metal plates) and will produce a corresponding proportional quantity of charge in the gas. By arranging that the plates are alternately biased at a positive and negative potential, the entire device operates like a large interleaved gas ion chamber. The total collected charge is then a measure of the total energy in the initial incident pulse of radiation.

Solid ionization chambers. Ionization chambers can be made where the initial ionization occurs, not in gases, but in suitable liquids or solids. In fact, the discovery of extremely successful solid-state ionization detectors in the early 1960s temporarily diverted interest from further developments of gas-filled chambers.

In the solid-state ionization chamber (or solid-state detector) the gas filling is replaced by a large single crystal of suitably chosen solid material. In this case the incident radiation creates electron-hole pairs in the crystal, and this constitutes the signal charge. In practice, it has been found that only very few materials can be produced with a sufficiently high degree of crystalline perfection to allow this signal charge to be swept out of the crystal and collected. Although many attempts were made in this direction in the 1940s in crystal counters, using such materials as AgCl, CdS, and diamond, these were all failures due to the crystals not having adequate carrier transport properties. In the late 1950s, however, new attempts were made in this direction using single crystals of the semiconductors silicon and germanium. These were highly successful and led to detectors that revolutionized low-energy nuclear spectroscopy.

There are two important differences between solid and gas-filled ionization chambers. First, it takes much less energy to create an electron-hole pair in a solid than it does to ionize gas atoms. Hence, the intrinsic energy resolution obtainable with solid-state detectors is better than with gas counters. Gamma-ray detectors with resolutions better than 180 eV are commercially available. Second, in the case of solid semiconductors, the positive charge is carried by electron holes whose mobilities are similar to those of electrons. Hence, both the electrons and holes are rapidly swept away by the electric field and, as a result, no Frisch grid is needed to electrically shield the anode from the image charge effects of slow-moving positive ions as in the case of gas- or liquid-filled ionization chambers. *See* CRYSTAL COUNTER; GAMMA-RAY DETECTORS; SEMICONDUCTOR.

Liquid ionization chambers. The use of a liquid in an ionization chamber combines many of the advantages of both solid and gas-filled ionization chambers; most importantly, such devices have the flexibility in design of gas chambers with the high density of solid chambers. The high density is especially important for highly penetrating particles such as gamma rays. Unfortunately, until the 1970s the difficulties of obtaining suitable high-purity liquids effectively stopped development of these detectors. During the 1970s, however, a number of groups built liquid argon ionization chambers and demonstrated their feasibility. A Frisch grid liquid argon chamber achieved a resolution of 34 keV (FWHM).

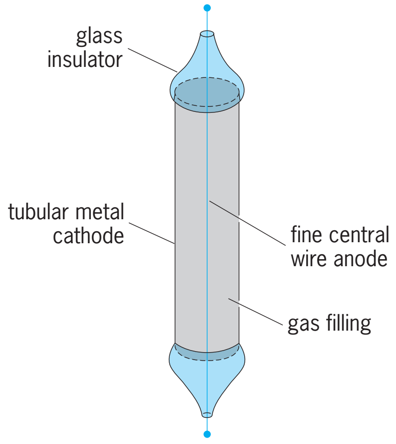


Fig. 7. Basic form of a simple single-wire gas proportional counter.

Proportional counters. If the electric field is increased beyond a certain point in a gas ionization chamber, a situation is reached in which the free electrons are able to create additional electron-ion pairs by collisions with neutral gas atoms. For this to occur, the electric field must be sufficiently high so that between collisions an electron can pick up an energy that exceeds the ionization potential of the neutral gas atoms. Under these circumstances gas multiplication, or avalanche gain, occurs, thereby providing additional signal charge from the detector.

A variety of electrode structures have been employed to provide proportional gas gain of this type. The most widely used is shown in **Fig. 7**. Here a fine central wire acts as the anode, and the avalanche gain takes place in the high field region immediately surrounding this wire. In practice, under suitable circumstances, it is possible to operate at gas gains of up to approximately 10^6.

The gas gain is a function of the bias voltage applied to the proportional counter and takes the general form shown in **Fig. 8**.

Similar avalanche multiplication effects can occur in semiconductor junction detectors, although there the situation is less favorable, and such devices have not found very widespread use except as optical detectors. *See* JUNCTION DETECTOR.

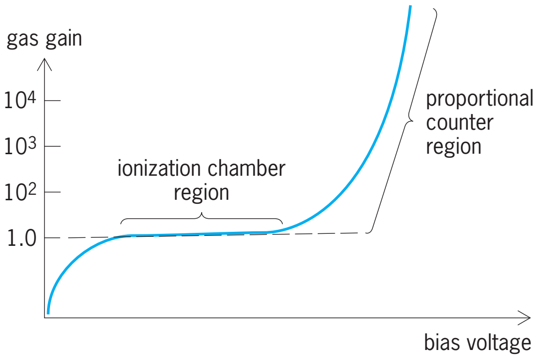


Fig. 8. Plot of gas gain versus applied voltage for a gas-filled radiation detector.

The large gas gains realizable with proportional counters have made them extremely useful for research applications involving very low-energy radiation. In addition, their flexibility in terms of geometry has made it possible to construct large-area detectors, of the order of 10 ft^2 (1 m^2), suitable for use as x-ray detectors in space. Essentially all that has been learned to date regarding x-ray astronomy has involved the use of such detectors aboard space vehicles. *See* X-RAY TELESCOPE.

Further exceedingly useful applications of gas proportional counters involve their use as position-sensitive detectors. In Fig. 7, for example, if the anode wire is grounded at both ends, then the signal charge generated at a point will split and flow to ground in the ratio of the resistance of the center wire between the point of origin and the two ends of the wire. This device therefore comprises a one-dimensional position-sensitive detector. Such devices are widely used as focal plane detectors in magnetic spectrographs. Similar position-sensitive operation can be obtained by taking account of the rise time of the signals seen at each end of the wire. Further extension of such methods allows two-dimensional detectors to be produced, a wide variety of which are under investigation for medical and other imaging uses.

The relatively large signals obtainable from gas proportional counters simplifies the requirements of the subsequent amplifiers and signal handling systems. This has made it economically feasible to employ very large arrays, on the order of thousands, of such devices in multidimensional arrays in high-energy physics experiments. By exploiting refinements of technique, it has proved possible to locate the tracks of charged particles to within a fraction of a millimeter in distances measured in meters. Such proportional counter arrays can operate at megahertz counting rates since they do not exhibit the long dead-time effects associated with spark chambers.

Geiger counters. If the bias voltage across a proportional counter is increased sufficiently, the device enters a new mode of operation in which the gas gain is no longer proportional to the initial signal charge but saturates at a very large, and constant, value. This provides a very economical method of generating signals so large that they need no subsequent amplification. *See* GEIGER-MÜLLER COUNTER.

The most widespread use of Geiger counters continues to be in radiation monitoring, where their large output signals simplify the readout problem. They have also found extensive use in cosmic-ray research, where again their large signals have made it feasible to use arrays of substantial numbers of detectors without excessive expenditures on signal-processing electronics. *See* PARTICLE DETECTOR.

William A. Lanford

Bibliography. D. A. Bromley (ed.), *Detectors in Nuclear Science*, 1979; C. F. Delaney and E. C. Finch, *Radiation Detectors: Physical Principles and Applications*, 1992; K. Kleinknecht, *Detectors for Particle Radiation*, 2d ed., 1999; G. F. Knoll, *Radiation Detection and Measurement*, 3d ed., 1999.

Ionization potential

The potential difference through which a bound electron must be raised to free it from the atom or molecule to which it is attached. In particular, the ionization potential is the difference in potential between the initial state, in which the electron is bound, and the final state, in which it is at rest at infinity.

The concept of ionization potential is closely associated with the Bohr theory of the atom. Although the simple theory is applicable only to hydrogenlike atoms, the picture furnished by it conveys the idea quite well. In this theory, the allowed energy levels for the electron are given by the equation below,

$$E_n = -\frac{k}{n^2} \qquad n = 1, 2, 3, \ldots$$

where E_n is the energy of the state described by n. The constant k is about 13.6 eV for atomic hydrogen. The energy approaches zero as n becomes infinite. Thus zero energy is associated with the free electron. On the other hand, the most tightly bound case is given by setting n equal to unity. By the definition given above, the ionization potential for the most tightly bound, or ground, state is then 13.6 eV. The ionization potential for any excited state is obtained by evaluating E_n for the particular value of n associated with that state. For a further discussion of the energy levels of an atom *see* ATOMIC STRUCTURE AND SPECTRA; ELECTRONVOLT.

The ionization potential for the removal of an electron from a neutral atom other than hydrogen is more correctly designated as the first ionization potential. The potential associated with the removal of a second electron from a single ionized atom or molecule is then the second ionization potential, and so on.

Ionization potentials may be measured in a number of ways. The most accurate measurement is obtained from spectroscopic methods. The transitions between energy states are accompanied by the emission or absorption of radiation. The wavelength of this radiation is a measure of the energy difference. The particular transitions that have a common final energy state are called a series. The series limit represents the transition from the free electron state to the particular state common to the series. The energy associated with the series limit transition is the ionization energy.

Another method of measuring ionization potentials is by electron impact. Here the minimum energy needed for a free electron to ionize in a collision is determined. The accuracy of this type of measurement cannot approach that of the spectroscopic method. *See* ELECTRON CONFIGURATION; ELECTRONEGATIVITY; STRUCTURAL CHEMISTRY. **Glenn H. Miller**

Bibliography. H. C. Ohanian, *Modern Physics*, 2d ed., 1995; J. W. Rohlf, *Modern Physics from A to Z*, 1994; T. R. Sandin, *Essentials of Modern Physics*, 1989.

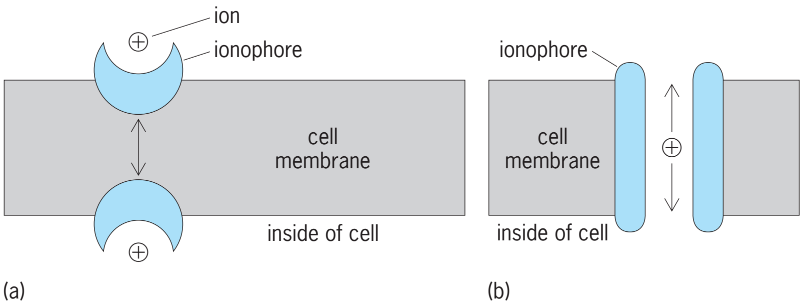
Ionophore

A substance that can transfer ions from a hydrophilic medium such as water into a hydrophobic medium, such as hexane or a biological membrane, where the ions typically would not be soluble, also known as an ion carrier. The ions transferred are usually metal ions [for example, lithium (Li^+), sodium (Na^+), potassium (K^+), magnesium (Mg^{2+}), or calcium (Ca^{2+})]; but there are ionophores that promote the transfer of other ions, such as ammonium ion (NH_4^+) or amines of biological interest. *See* ION.

Ionophores were discovered in the early 1960s, when it was found that certain antibiotics, whose mechanism of action was unknown at the time, depended on the presence of ions such as potassium for their biological activity. Mitochondria isolated from cells would swell when placed in an aqueous medium containing small amounts of potassium ions and the antibiotic valinomycin. Measurement of the pH and potassium-ion concentration of the medium revealed that the potassium concentration decreased while the hydrogen concentration increased correspondingly. Thus, valinomycin was catalyzing the exchange of potassium ions in the medium for hydrogen ions present in the mitochondria. The swelling of the mitochondria was found to be due to an increase in their water content, which was necessary to rehydrate the potassium ions. The effect was reasonably specific for potassium ion, as smaller ions such as sodium and lithium were without effect; and, if either valinomycin or the potassium ions were independently eliminated from the experiment, no swelling was observed. Thus valinomycin had promoted the transfer of potassium ion across a hydrophobic cell membrane.

Mechanism of transfer. There are two different mechanisms by which ionophores promote the transfer of ions across hydrophobic barriers: ion-ionophore complex formation and ion channel formation (see **illus.**).

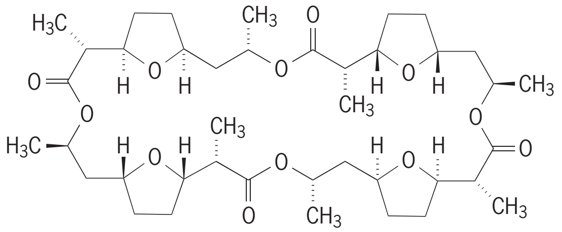
In complex formation, the ion forms a coordination complex with the ionophore in which there is a well-defined ratio (typically 1:1) of ion to ionophore. In these complexes the ionophore wraps around the ion so that the ion exists in the polar interior of the complex while the exterior is predominantly hydrophobic in character, and as such is



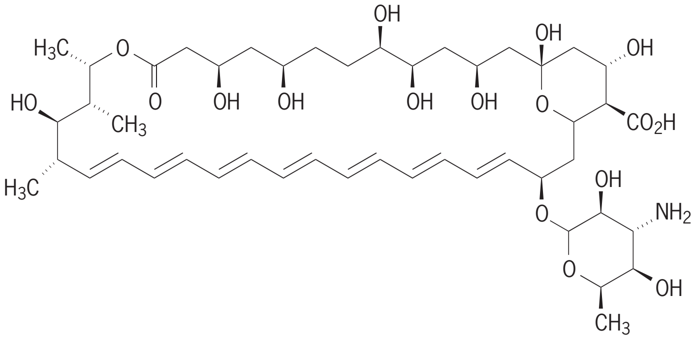
Two mechanisms by which ionophores promote ion transfer across cell membranes: (*a*) ion-ionophore complex and (*b*) ion-channel-formation ion transfer.

soluble in nonpolar media. In these complexes the ion is coordinated by oxygen atoms present in the ionophore molecule through ion-dipole interactions. The ionophore molecule essentially acts as the solvent for the ion, replacing the aqueous solvation shell that normally surrounds the ion. An example of an ionophore that forms ion-ionophore complexes is nonactin (**1**). *See* COORDINATION COMPLEXES.

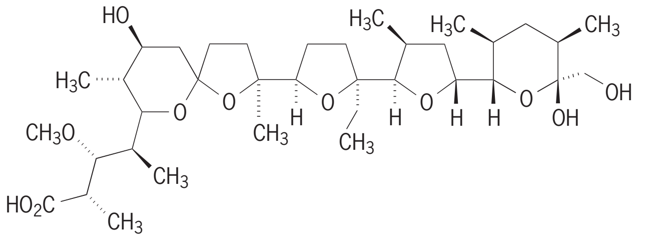
Ionophores that act via ion channel formation are found in biological environments. In this type of ionophore the molecule forms a polar channel in an otherwise nonpolar cell membrane, allowing passage of small ions either into or out of the cell. An example of an ion-channel-forming ionophore is amphotericin-B (**2**).



(1)



(2)



(3)

Types. Ionophores were discovered as metabolites of microorganisms, but as interest in these compounds has increased, compounds possessing ionophoric activity have been designed and synthesized by chemists.

Naturally occurring compounds. There are several different types of naturally occurring ionophores. These have been classified into four classes, each of which has antibiotic activity.

1. Peptide ionophores are exemplified by the gramicidins, a family of polypeptide-containing antibiotics. These compounds are thought to form ion channels via the dimerization of two gramicidin molecules to form a helical conformation in a cell membrane. Ions may then pass down the polar interior of the helix.

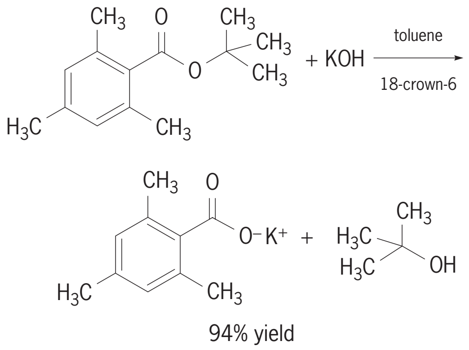
2. Valinomycin is an example of the cyclic depsipeptide class of ionophores. The molecule is a 36-membered macrocycle that is made up of a repetition of valine, hydroxyvaleric acid, and lactic acid residues. The structure of valinomycin complexed to potassium ion has been determined by x-ray crystallography. In this complex the conformation of the molecule is such that the potassium ion resides in the center of the complex coordinated to six carbonyl oxygens that are arranged in an octahedral configuration about the metal ion.

3. Nonactin (**1**) is a representative of the macrotetrolide class of ionophores. In this macrocycle the tetrahydrofuranyl oxygen atoms and two of the four lactone carbonyl oxygen atoms are responsible for chelation of metal ions. *See* CHELATION.

4. The polyether ionophores form a large family of structurally complex compounds. Monensin (**3**) is the best-known member of this class and was the first polyether ionophore in which the structure was determined by x-ray crystallography. Members of this class of ionophores have in common the presence of a carboxylic acid functional group (COOH) on one end of the molecule and a hydroxyl group (OH) on the other end. In solution these groups link to each other with hydrogen bonds, folding the molecule into a cyclic array. The metal ion sits in the center of the complex and is coordinated to the ether oxygen atoms of the tetrahydrofuran and tetrahydropyran rings.

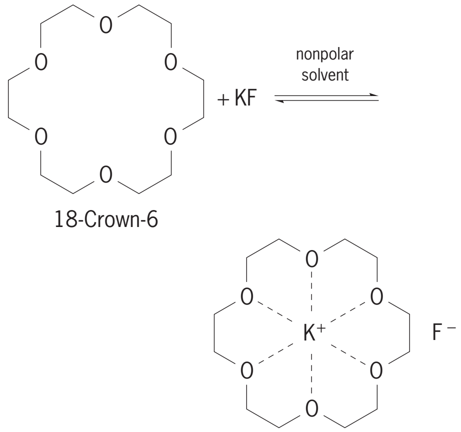
Synthetic compounds. In addition to the naturally occurring ionophores, a large number of compounds possessing ionophoric activity have been synthesized. The best-known synthetic ionophores are the crown ethers, which are cyclic polyethers containing the repeating units of (—O—CH_2—CH_2—). The name crown ethers is due to the resemblance of the molecule's conformation to a king's crown in which the oxygen atoms form the points of the crown. Simple crown ethers are named by the size of the ring and the number of oxygen atoms contained. For example, the name 18-crown-6 refers to an 18-membered ring containing 6 ether oxygen atoms. As with other ionophores, crown ether complex ions are coordinated through a series of ion-dipole interactions occurring between the ion and the ether oxygen atoms present in the ionophore. *See* MACROCYCLIC COMPOUND.

Synthetic ionophores such as the crown ethers have proven to be useful reagents in synthetic organic chemistry. These reagents can solubilize inorganic salts, for example, potassium hydroxide (KOH), in nonpolar solvents such as toluene, in which they normally would be insoluble. This capability allows reactions involving these salts to be successfully conducted in the absence of water. An example of the utility of these reagents is the very efficient hydrolysis of the ester group in *tert*-butyl mesitoate [reaction (1)], which is readily accomplished by potassium hydroxide in toluene in the presence of a potassium-complexing crown ether. No reaction occurs under identical conditions in the absence of the crown ether.



(1)

Many inorganic reagents (generically denoted as M^+X^-) are found to be much more reactive when their reactions are conducted in nonpolar solvents containing crown ethers. The increased reactivity is due to the poor solvation of the anion (X^-). Potassium fluoride (KF) provides an example of this effect. Normally, in aqueous solution the fluoride ion (F^-) is heavily solvated by water molecules and is not very reactive. However, when potassium fluoride is dissolved in nonpolar solvents by employing a crown ether as the cation-solvating agent [reaction (2)], the unsolvated fluoride ion becomes a very powerful nucleophile. *See* ELECTROPHILIC AND NUCLEOPHILIC REAGENTS.



(2)

Biological activity antibiotics. The biological activity of ionophore antibiotics is due to their ability to disrupt the flow of ions either into or out of cells. Under normal conditions cells have a high internal concentration of potassium ions but a low concentration of sodium ions. The concentration of ions in the extracellular medium is just the reverse, high in sodium ions but low in potassium ions. This imbalance, which is necessary for normal cell function, is maintained by the presence of a specific transport protein (sodium-potassium ATPase) present in the cell membrane that pumps sodium ions out of the cell in exchange for potassium ions. Ionophore antibiotics possess the ability to disrupt this ionic imbalance by allowing ions to penetrate the cell membrane as ion-ionophore complexes or via the formation of ion channels. Gram-positive bacteria appear to be particularly sensitive to the effect of ionophores perturbing normal ion transport. *See* ION TRANSPORT.

The pharmacological use of many of the ionophore antibiotics in humans is very restricted because of their high toxicity and multitude of physiological effects. *See* ANTIBIOTIC. Chris A. Veale

Bibliography. M. Dobler, *Ionophores and Their Structures*, 1981.

Ionosphere

The part of the upper atmosphere that is sufficiently ionized that the concentration of free electrons affects the propagation of radio waves. Existence of the ionosphere was suggested simultaneously in 1902 by O. Heaviside in England and A. E. Kennelly in the United States to explain the transatlantic radio communication that was demonstrated the previous year by G. Marconi; and for many years it was commonly referred to as the Kennelly-Heaviside layer. The existence of the ionosphere as an electronically conducting region had been postulated earlier by B. Steward to explain the daily variations in the geomagnetic field. *See* IONIZATION; RADIO-WAVE PROPAGATION.

The ionosphere has been extensively explored by rockets, satellites, and remote-sensing tools. The earliest technique involved the ionosonde, which utilizes a pulsed transmitter to send radio signals vertically upward while slowly sweeping the radio frequency. The ionosphere is a type of plasma, and such a system has a free oscillation mode. When the frequency of the signal transmitted by the ionosonde reaches a height where it equals the free oscillation frequency of the plasma, it reflects to Earth. Pulses reflected by the ionosphere are received at the transmitter and recorded; the elapsed time between pulse transmission and reception can be measured, and by using the speed of light it can be converted to an apparent distance to the point of reflection. The plasma density content can be found in this way by using an ionosonde, but no information is obtained above the point where the plasma density is maximum, so properties of the ionosphere above about 180 mi (300 km) altitude were impossible to determine. Major breakthroughs occurred in space science with the development of the artificial space satellite and sounding rockets. *See* METEOROLOGICAL ROCKET; OSCILLATION; PLASMA (PHYSICS).

Another tool for ionospheric research is the incoherent scatter radar. Powerful radio sources, such a radar, can probe out to many thousand kilometers' altitude and detect the weak signal scattered back to Earth by thermal fluctuations in the plasma. The first two of these instruments were built in 1960 near Arecibo, Puerto Rico, and Jicamarca, Peru; both are roughly 1000 ft (300 m) across. The Arecibo system is so sensitive that it is used for radio astronomy as well as ionospheric and atmospheric science. Eight radars of this type form an observatory network stretching north-south in the Americas from Peru to Greenland and east-west from Japan to Europe. Using the Doppler technique, much like the radar in a police car, such a system can be used to measure the velocity of the plasma as well as its temperature and density. *See* DOPPLER RADAR; RADAR.

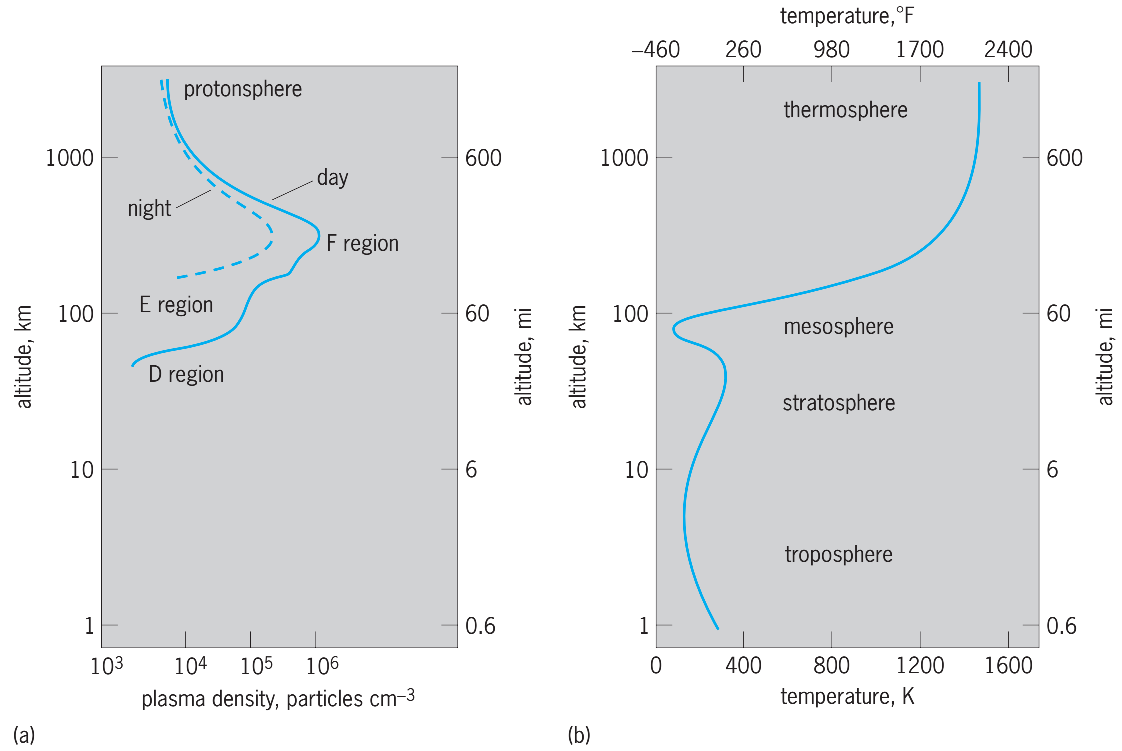
Regions. The ionosphere is highly structured in the vertical direction. It was first thought that discrete layers were involved, referred to as the D, E, F_1, and F_2 layers; however, the layers actually merge with one another to such an extent that they are now referred to as regions rather than layers. The very high temperatures in the Earth's upper atmosphere are colocated with the upper ionosphere (see **illus.**), since both are related to the effect of x-rays from the Sun. That is, the x-rays both ionize and heat the very uppermost portion of the Earth's atmosphere.

The ionosphere also displays striking geographic and temporal variations; the latter include regular diurnal, seasonal, and sunspot-cycle components and irregular day-to-day components that are associated mainly with variations in solar activity and atmospheric motion. *See* SUN.

D region. The D region is the lowest ionospheric region, extending approximately from 35 to 50 mi (60 to 85 km). The upper portion is caused mainly by the ionization of nitric oxide by Lyman-alpha radiation in sunlight, and the lower portion is mainly due to ionization by cosmic radiation. The daytime electron concentrations are about 10^8–10^9 per cubic meter. The region virtually disappears at night even though the cosmic radiation continues, because attachment of electrons to molecules (that is, formation of negative ions) quickly removes free electrons. This effect is suppressed during the daytime by photo detachment. The D region is the only ionospheric region in which negative ions are thought to be significant. *See* COSMIC RAYS.

The collision frequency for electrons with heavier particles in the D region is relatively high, which causes absorption of energy from radio signals traveling through the region. This severely limits radio propagation and is responsible for the very limited daytime range for stations in the AM broadcast band; FM radio and television signals are at such a high frequency that they pass directly through the ionosphere with no reflection possible. At night, when the D region disappears, lower-frequency radio signals can reflect many times from the ionosphere and the Earth, leading to the vast propagation distances used by amateur radio operators for communication. *See* RADIO-WAVE PROPAGATION.

The D region is located within the Earth's atmospheric region known as the mesosphere (see illus.). The coldest temperatures on Earth and the highest clouds occur here. The latter occur in the summer polar regions where temperatures below −280°F (100 K) have been recorded. These clouds are called noctilucent since they can be seen only at twilight, illuminated by sunlight at their height near 48 mi (80 km) and viewed from the dark surface of the Earth. The clouds are composed of ice; they are charged electrically and form a different type of



Atmospheric profiles. (*a*) Ionosphere, nighttime plasma density. (*b*) Thermal structure of the atmosphere. (*After M. C. Kelley, The Earth's Ionosphere: Plasma Physics and Electrodynamics, Academic Press, 1989*)

material known as dusty or icy plasma. *See* MESOSPHERE.

E region. Soft x-rays and the more penetrating portions of the extreme ultraviolet radiation from the Sun are absorbed in the altitude region from 50 to 85 mi 85 to (140 km), where they cause daytime electron concentrations of the order of 10^{11} per cubic meter in the E region (see illus.). This is the region from which ionospheric reflections were first identified. The principal ions have been observed to be the oxygen ions O_2^+ and O^+ and the oxynitrogen ion NO^+, the last presumably being formed in chemical reactions involving the primary oxygen ions. The soft x-rays that are principally responsible for the formation of the E region must also produce nitrogen ions (N_2^+); however, these are not observed because they are removed very rapidly by chemical reactions. When the Sun sets, these molecular ions react very rapidly with equal numbers of electrons, yielding oxygen and nitrogen atoms. These reactions are so fast that the E region virtually disappears at night unless there is an aurora present. This explains the low level of nighttime plasma density in the ionospheric profile. (illus. *a*).

The Earth is continually bombarded by meteors which burn up and become ionized when they hit the top of the atmosphere. This leaves a layer of dust, metallic atoms, and metallic ions in the E region. Ions such as iron (Fe^+), sodium (Na^+), and silicon (Si^+) do not recombine quickly because the atom cannot break apart. The result is the sporadic occurrence of long-lived clouds of metallic ion layers in the E region. The fading in and out of distant AM radio signals is due to their reflection to Earth from these sporadic E layers. *See* METEOR.

The E region supports the largest electric currents in the near space regions of the Earth. Currents as high as 10^6 A flow in the polar E region during an intense aurora. In fact, the light associated with the most beautiful of the auroral displays is generated in this region of the ionosphere and atmosphere. The electrical power that is transferred to the atmosphere in such an event is about 100 GW.

F region. The solar extreme-ultraviolet radiations produce the F region, above 85 mi (140 km). This radiation is most strongly absorbed near 95 mi (160 km). Above the region of maximum photoionization, the rate for the loss of electrons and ions decreases with altitude more rapidly than does the source of ionization, so that electron concentrations increase with altitude. This decrease in loss rate occurs because O^+ ions recombine directly with electrons only very slowly in a two-body radiative-recombination process. Another loss process predominates instead, an ion-atom interchange reaction [reaction (1)], followed by a dissociative recombination [reaction (2)]. The loss rate is controlled by the

$$O^+ + N_2 \rightarrow NO^+ + N \qquad (1)$$

$$NO^+ + e^- \rightarrow N + O \qquad (2)$$

ion-atom interchange reaction. Since the N_2 concentration decreases with altitude, the concentration of O^+ increases with altitude above the region where it is actually produced.

The increase in O^+ and electron concentration with altitude finally stops, because the atmospheric density becomes so low that there is a rapid downward flow of ions and electrons in the gravitational field. The region of maximum ionization concentration normally occurs near 180 mi (300 km), and it is known as the F peak. Most of the ionization that occurs above the peak is lost by flow into the dense atmosphere below, where the ionization can be lost by the ion-atom interchange reaction followed by dissociative recombination [reactions (1) and (2)].

The peak daytime plasma concentration in the F region is in the vicinity of 10^{12} ions per cubic meter. Above the peak, the distribution of ionization is in diffusive equilibrium in the gravitational field. The colocated atmospheric region (the thermosphere) has this property as well. Unlike the atmosphere below 60 mi (100 km), which is thoroughly mixed, the various atoms and ions can separate by their mass, the lighter particles extending to very high altitudes. Hydrogen and helium can, in fact, escape the Earth's pull.

The Earth's neutral thermosphere reaches temperatures of well over 1500°F (1188 K) and exchanges the associated energy with the plasma in the ionosphere. This high temperature is due to the absorption of the Sun's deadly x-rays and extreme-ultraviolet light, which also creates the ionized gas. Just as the ozone layer absorbs ultraviolet light in the upper stratosphere and is colocated with a temperature increase, the thermospheric temperature rise marks an important filter of these high-energy solar photons.

The thermosphere is so tenuous that the temperature at a given point responds quickly to solar heat input. The result is a huge temperature difference between night and day, as much as 900°F (500 K). In response, intense winds exceeding 450 mi/h (200 m/s) blow back and forth from high to low temperature zones. As the Earth rotates under this pattern, an observer sees a strong component at a 24-h period. This is called a solar thermal tide in analogy to the 12-h periodicity exhibited by the Earth's oceans responding to the gravitational attraction of the Moon. *See* MOON.

The ionosphere responds to these neutral tidal motions, but with a behavior that is constrained by the Earth's magnetic field and the electric currents that are generated in the conducting fluid. These wind-driven currents also create electric fields. The combined action of electric, gravitational, and magnetic fields, along with the neutral wind, controls the ionospheric material. *See* GEOMAGNETISM; THERMOSPHERE.

Heliosphere and protonosphere. Helium and atomic hydrogen are important constituents of the upper atmosphere, and their ions are also important at levels above 300 mi (500 km) or so. These gases become important, and finally predominant, constituents of the upper atmosphere because of their low mass. In diffusive equilibrium, each neutral gas is distributed in the gravitational field just as if the other gases were not present, and the lighter ones therefore finally come to predominate over the heavier ones above some sufficiently high altitude. The terms heliosphere and protonosphere are sometimes used to designate the regions in which helium and hydrogen ions respectively are predominant. *See* HELIUM; HYDROGEN.

At midlatitudes these light ions dominate above about 600 mi (1000 km). Also above this height the magnetic field geometry becomes crucial, and a toruslike region called the plasmasphere forms; it is full of dense cool plasma. The torus extends to about four earth radii (11,4000 mi or 19,000 km above the surface) and, following the curved magnetic field lines, this boundary touches the top of the atmosphere at all latitudes below about 60° magnetic latitude (southern Canada in the American Sector). This torus can be considered an extension of the ionosphere, since it is filled from below by the ionosphere during the day and empties back into the ionosphere during the night. The plasmasphere is relatively stable, since it rotates with the planet just as the Earth's gaseous neutral atmosphere does. At the higher latitudes the ionosphere is put into motion by the solar wind and no longer rotates with the planet. *See* SOLAR WIND.

Electrodynamics. In an electrical generator a conducting material is moved across a magnetic field, and the resulting magnetic force drives a current. If there is no external connection, though, the electric charge piles up at the ends of the conductor and an electric field is generated. The electric field attains a magnitude that balances the magnetic forces and cancels the current. If a light bulb is attached to the generator, a small current will flow. In this case an electric field will still exist, but it will be smaller than in the case with no external connection.

The same effect occurs when the Earth's winds blow across its magnetic field. A current is generated by the magnetic force. But when boundaries are encountered, an electric field will be created. These electric fields greatly affect the dynamics of the ionosphere. This wind-driven source of electric fields dominates the electrodynamics of the ionosphere below about 60° latitude.

At higher latitudes the Earth's atmosphere loses control of the plasma, and the plasma is subject to strong electrical forces induced by the flow of the solar wind past the Earth. Up to 200,000 V can be impressed across the Earth's ionosphere when the Sun's magnetic field is aligned opposite to the Earth's. The high-latitude ionosphere is then put into a high-speed circulation characterized by two vortices in each hemisphere. The resulting electrical force can, through collisional forces, create a similar circulation in the upper-atmosphere neutral air. *See* SOLAR MAGNETIC FIELD.

Tremendous variations occur in the ionosphere at high latitudes because of the dynamical effects of these electrical forces and because of the additional sources of plasma production. The most notable is the visual aurora, one of the most spectacular natural sights. Views of the Earth from both scientific and defense satellites have revealed the extent and beauty of the aurora against the background of cities. *See* MAGNETOSPHERE; MILITARY SATELLITES.

The aurora has a poleward and equatorward limit during times of magnetic storms. A resident of the arctic regions of the Northern Hemisphere see the "northern" lights in their southern sky, for example. The aurora forms two rings around the poles of the Earth. The size of the rings waxes and wanes while wavelike disturbances propagate along its extent. *See* AURORA; UPPER-ATMOSPHERE DYNAMICS.

Space plasma physics. The ionosphere is subject to the same disruptions that plague attempts to control nuclear fusion by magnetic confinement. By using the principles of plasma physics, it is possible to explain many of the processes that were known since the early twentieth century to seriously affect radio communication systems. The culprits are generically referred to as plasma instabilities, and they occur when stored energy is rapidly released. Typically such a process is associated with localized turbulent zones and an irregular ionosphere, similar to severe weather in the lower atmosphere.

At low latitudes the most dominant process is the Rayleigh-Taylor instability, in which gravitational energy is released after sunset. Huge convective "thunderstorms" arise that race around the Earth, following the sunset terminator. Intense currents in the lower ionosphere also create large-amplitude plasma sound waves via the Farley-Buneman instability all over the Earth, but with two intense regions: one near the magnetic equator and one at high latitudes where the magnetic field is nearly vertical.

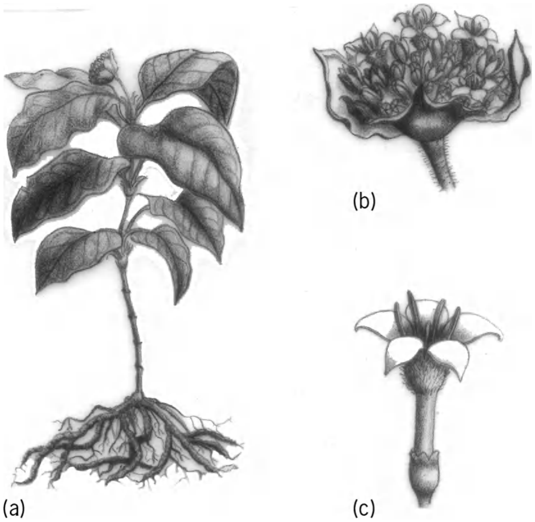
Some of these instabilities are induced by neutral atmospheric waves that are generated in the dense lower atmosphere and propagate upward. These gravity or buoyancy waves have wavelengths of hundreds of kilometers and are created, among other things, by weather fronts, mountain ranges, earthquakes, nuclear explosions, and aurora. Even without their role as a source of plasma instabilities, these waves push the ionosphere around, creating both wavelike horizontal undulations and intense layers in the metallic materials from meteors, the sporadic E layers. *See* ATMOSPHERE.

Michael C. Kelley; Francis S. Johnson

Bibliography. K. Davies, *Ionospheric Radio*, 1990; A. Giraud and M. Petit, *Ionospheric Techniques and Phenomena*, 1978; A. Giraud and M. Petit, *Physics of the Earth's Ionosphere*, 1978; J. K. Hargreaves, *The Solar-Terrestrial Environment*, 1992; M. C. Kelley, *The Earth's Ionosphere: Electrodynamics and Plasma Physics*, 1989; J. A. Ratcliffe, *An Introduction to the Ionosphere and Magnetosphere*, 1972.

Ipecac

A low, perennial shrub or half-shrub of the tropical forest in Brazil and Colombia. Several species are used, but the dried roots and rhizomes of *Cephaelis*



Ipecac (*Cephaelis ipecacuanha*). (*a*) Entire plant. (*b*) Inflorescence. (*c*) Flower.

ipecacuanha (Rubiaceae) constitute the material recognized as the official drug ipecac (see **illus.**). Medically the principal use is as an emetic and an expectorant. *See* RUBIALES.

P. D. Strausbaugh; Earl L. Core

Iridium

A chemical element, Ir, atomic number 77, relative atomic weight 192.22. Iridium is a transition metal and shares similarities with rhodium as well as the other platinum metals, including palladium, platinum, ruthenium, and osmium. The atom in the gas phase has the electronic configuration $1s^2$, $2s^2$, $2p^6$, $3s^2$, $3p^6$, $3d^{10}$, $4s^2$, $4p^6$ $4d^{10}$, $4f^{14}$, $5s^2$, $5p^6$, $5d^7$, $6s^2$. The ionic radius for Ir^{3+} is 0.068 nanometer and its metallic radius is 0.1357 nm. Metallic iridium is slightly less dense than osmium, which is the densest of all the elements. *See* PERIODIC TABLE.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1 **H**																	2 **He**
3 **Li**	4 **Be**											5 **B**	6 **C**	7 **N**	8 **O**	9 **F**	10 **Ne**
11 **Na**	12 **Mg**											13 **Al**	14 **Si**	15 **P**	16 **S**	17 **Cl**	18 **Ar**
19 **K**	20 **Ca**	21 **Sc**	22 **Ti**	23 **V**	24 **Cr**	25 **Mn**	26 **Fe**	27 **Co**	28 **Ni**	29 **Cu**	30 **Zn**	31 **Ga**	32 **Ge**	33 **As**	34 **Se**	35 **Br**	36 **Kr**
37 **Rb**	38 **Sr**	39 **Y**	40 **Zr**	41 **Nb**	42 **Mo**	43 **Tc**	44 **Ru**	45 **Rh**	46 **Pd**	47 **Ag**	48 **Cd**	49 **In**	50 **Sn**	51 **Sb**	52 **Te**	53 **I**	54 **Xe**
55 **Cs**	56 **Ba**	71 **Lu**	72 **Hf**	73 **Ta**	74 **W**	75 **Re**	76 **Os**	77 **Ir**	78 **Pt**	79 **Au**	80 **Hg**	81 **Tl**	82 **Pb**	83 **Bi**	84 **Po**	85 **At**	86 **Rn**
87 **Fr**	88 **Ra**	103 **Lr**	104 **Rf**	105 **Db**	106 **Sg**	107 **Bh**	108 **Hs**	109 **Mt**	110 **Ds**	111 **Rg**	112	113					

lanthanide series	57 **La**	58 **Ce**	59 **Pr**	60 **Nd**	61 **Pm**	62 **Sm**	63 **Eu**	64 **Gd**	65 **Tb**	66 **Dy**	67 **Ho**	68 **Er**	69 **Tm**	70 **Yb**
actinide series	89 **Ac**	90 **Th**	91 **Pa**	92 **U**	93 **Np**	94 **Pu**	95 **Am**	96 **Cm**	97 **Bk**	98 **Cf**	99 **Es**	100 **Fm**	101 **Md**	102 **No**

The abundance of iridium in the Earth's crust is very low, 0.001 ppm. For mining purposes, it is generally found alloyed with osmium in materials known as osmiridium and iridiosmium, with iridium contents ranging from 25 to 75%.

Solid iridium is a silvery metal with considerable resistance to chemical attack. Upon atmospheric exposure the surface of the metal is covered with a relatively thick layer of iridium dioxide (IrO_2). Important physical properties of metallic iridium are given in the **table**.

Because of its scarcity and high cost, applications of iridium are severely limited. Although iridium metal and many of its complex compounds are good catalysts, no large-scale commercial application for these has been developed. In general, other platinum metals have superior catalytic properties. The high degree of thermal stability of elemental iridium and the stability it imparts to its alloys does give rise to those applications where it has found success. Particularly relevant are its high melting point (2443°C or 4429°F), its oxidation resistance, and the fact that it is the only metal with good mechanical properties that survives atmospheric exposure above 1600°C (2910°F). Iridium is alloyed with platinum to increase tensile strength, hardness, and corrosion resistance. However, the workability of these alloys is decreased. These alloys find use as electrodes for anodic oxidation, for containing and manipulating corrosive chemicals, for electrical contacts that are exposed to corrosive chemicals, and as primary standards for weight and length. Platinum-iridium alloys are used for electrodes in spark plugs that are unusually resistant to fouling by antiknock lead additives. Iridium-rhodium thermocouples are used for high-temperature applications, where they have unique stability. Very pure iridium crucibles are used for growing single crystals of gadolinium gallium garnet for computer memory devices and of yttrium aluminum garnet for solid-state lasers. The radioactive

Physical properties of iridium metal

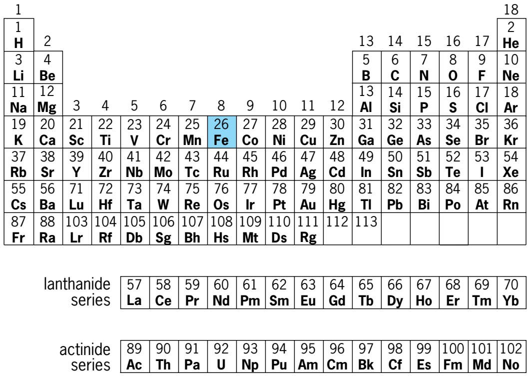
Property	Value
Crystal structure	Face-centered cubic
Lattice constant *a* at 25°C, nm	0.38394
Thermal neutron capture cross section, barns	440
Density at 25°C, g/cm^3	22.560
Melting point	2443°C (4429°F)
Boiling point	4500°C (8130°F)
Specific heat at 0°C, cal/g	0.0307
Thermal conductivity 0–100°C, cal cm/cm^2 s °C	0.35
Linear coefficient of thermal expansion 20–100°C, μin./in./°C	6.8
Electrical resistivity at 0°C, microhm-cm	4.71
Temperature coefficient of electrical resistance 0–100°C/°C	0.00427
Tensile strength (1000 lb/in.^{2})	
Soft	160–180
Hard	300–360
Young's modulus at 20°C	
lb/in.^{2}, static	75.0×10^6
lb/in.^{2}, dynamic	76.5×10^6
Hardness, diamond pyramid number	
Soft	200–240
Hard	600–700
ΔH_{fusion}, kJ/mol	26.4
$\Delta H_{vaporization}$, kJ/mol	612
ΔH_f monoatomic gas, kJ/mol	669
Electronegativity	2.2

isotope, ^{192}Ir, which is obtained synthetically from ^{191}Ir by irradiation of natural sources, has been used as a portable gamma source for radiographic studies in industry and medicine. *See* HIGH-TEMPERATURE MATERIALS; PLATINUM.

Bibliography. J. D. Atwood (ed.), *Comprehensive Organometallic Chemistry II: Cobalt, Rhodium and Iridium*, 1995; F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; A. Earnshaw and N. Greenwood, *Chemistry of the Elements*, 2d ed., Butterworth-Heinemann, 1997.

Iron

A chemical element, Fe, atomic number 26, and atomic weight 55.847. Iron is the fourth most abundant element in the crust of the Earth (5%). It is a malleable, tough, silver-gray, magnetic metal. It melts at 1540°C, boils at 2800°C, and has a density of 7.86 g/cm^3. The four stable, naturally occurring isotopes have masses of 54, 56, 57, and 58. The two main ores are hematite, Fe_2O_3, and limonite, $Fe_2O_3 \cdot 3H_2O$. Pyrites, FeS_2, and chromite, $Fe(CrO_2)_2$, are mined as ores for sulfur and chromium, respectively. Iron is found in many other minerals, and it occurs in groundwaters and in the red hemoglobin of blood. *See* PERIODIC TABLE.



1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1 H																	2 He
3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
11 Na	12 Mg											13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
55 Cs	56 Ba	71 Lu	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
87 Fr	88 Ra	103 Lr	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112	113					

lanthanide series	57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb
actinide series	89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No

The greatest use of iron is for structural steels; cast iron and wrought iron are made in quantity, also. Magnets, dyes (inks, blueprint paper, rouge pigments), and abrasives (rouge) are among the other uses of iron and iron compounds. *See* CAST IRON; IRON ALLOYS; IRON METALLURGY; STAINLESS STEEL; STEEL MANUFACTURE; WROUGHT IRON.

There are several allotropic forms of iron. Ferrite or α-iron is stable up to 760°C (1400°F). The change of β-iron involves primarily a loss of magnetic permeability because the lattice structure (body-centered cubic) is unchanged. The allotrope called γ-iron has the cubic close-packed arrangements of atoms and is stable from 910 to 1400°C (1670 to 2600°F). Little is known about δ-iron except that it is stable above 1400°C (2600°F) and has a lattice similar to that of α-iron.

The metal is a good reducing agent and, depending on conditions, can be oxidized to the 2+, 3+, or 6+ state. In most iron compounds, the ferrous ion, iron(II), or ferric ion, iron(III), is present as a distinct unit. Ferrous compounds are usually light yellow to dark green-brown in color; the hydrated ion, $Fe(H_2O)_6^{2+}$, which is found in many compounds and in solution, is light green. This ion has little tendency to form coordination complexes except with strong reagents such as cyanide ion, polyamines, and porphyrins. The ferric ion, because of its high charge (3+) and its small size, has a strong tendency to hold anions. The hydrated ion, $Fe(H_2O)_6^{3+}$, which is found in solution, combines with OH^-, F^-, Cl^-, CN^-, SCN^-, N_3^-, $C_2O_4^{2-}$, and other anions to form coordination complexes. *See* COORDINATION CHEMISTRY.

An interesting aspect of iron chemistry is the array of compounds with bonds to carbon. Cementite, Fe_3C, is a component of steel. The cyanide complexes of both ferrous and ferric iron are very stable and are not strongly magnetic in contradistinction to most iron coordination complexes. The cyanide complexes form colored salts. *See* TRANSITION ELEMENTS.

John O. Edwards

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; N. N. Greenwood and A. Earnshaw, *Chemistry of the Elements*, 2d ed., 1997; J. Silver (ed.), *Chemistry of Iron*, 1993.

Iron alloys

Solid solutions of metals, one metal being iron. A great number of commercial alloys have iron as an intentional constituent. Iron is the major constituent of wrought and cast iron and wrought and cast steel. Alloyed with usually large amounts of silicon, manganese, chromium, vanadium, molybdenum, niobium (columbium), selenium, titanium, phosphorus, or other elements, singly or sometimes in combination, iron forms the large group of materials known as ferroalloys that are important as addition agents in steelmaking. Iron is also a major constituent of many special-purpose alloys developed to have exceptional characteristics with respect to magnetic properties, electrical resistance, heat resistance, corrosion resistance, and thermal expansion. **Table 1** lists some of these alloys. *See* ALLOY; FERROALLOY; STEEL.

Because of the enormous number of commercially available materials, this article is limited to the better-known types of alloys. Emphasis is on special-purpose alloys; practically all of these contain relatively large amounts of an alloying element or elements referred to in the classification. Alloys containing less than 50% iron are excluded, with a few exceptions.

Iron-aluminum alloys. Although pure iron has ideal magnetic properties in many ways, its low electrical resistivity makes it unsuitable for use in alternating-current (ac) magnetic circuits. Addition of aluminum in fairly large amounts increases the electrical

TABLE 1. Some typical composition percent ranges of iron alloys classified by important uses*

Type	Fe	C	Mn	Si	Cr	Ni	Co	W	Mo	Al	Cu	Ti
Heat-resistant alloy castings	Bal.	0.30–0.50	—	1–2	8–30	0–7						
	Bal.	0.20–0.75	—	2–2.5	10–30	8–41						
Heat-resistant cast irons	Bal.	1.8–3.0	0.3–1.5	0.5–2.5	15–35	5 max						
	Bal.	1.8–3.0	0.4–1.5	1.0–2.75	1.75–5.5	14–30	—	—	1	—	7	
Corrosion-resistant alloy castings	Bal.	0.15–0.50	1 max	1	11.5–30	0–4	—	0.5 max				
	Bal.	0.03–0.20	1.5 max	1.5–2.0	18–27	8–31						
Corrosion-resistant cast irons	Bal.	1.2–4.0	0.3–1.5	0.5–3.0	12–35	5 max	—	—	4 max		3 max	
	Bal.	1.8–3.0	0.4–1.5	1.0–2.75	1.75–5.5	14–32	—	—	1 max		7 max	
Magnetically soft materials	Bal.	—	—	0.5–4.5								
	Bal.	—	—	—	—	—	—	—	3.5	16		
	Bal.	—	—	—	—	—	—	—	—	16		
	Bal.	—	—	—	—	—	—	—	—	12		
Permanent-magnet materials	Bal.	—	—	—	—	—	12	—	17			
	Bal.	—	—	—	—	—	12	—	20			
	Bal.	—	—	—	—	20	5	—	—	12		
	Bal.	—	—	—	—	17	12.5	—	—	10	6	
	Bal.	—	—	—	—	25	–	—	—	12		
	Bal.	—	—	—	—	28	5	—	—	12		
	Bal.	—	—	—	—	14	24	—	—	8	3	
	Bal.	—	—	—	—	15	24	—	—	8	3	1.25
Low-expansion alloys	Bal.	—	0.15	0.33	—	36						
	Bal.	—	0.24	0.03	—	42						
	61–53	0.5–2.0	0.5–2.0	0.5–2.0	4–5	33–35	—	1–3				

*This table does not include any AISI standard carbon steels, alloy steels, or stainless and heat-resistant steels or plain or alloy cast iron for ordinary engineering uses; it includes only alloys containing at least 50% iron, with a few exceptions. Bal. = balance percent of composition.

resistivity of iron, making the resulting alloys useful in such circuits.

Three commercial iron-aluminum alloys having moderately high permeability at low field strength and high electrical resistance nominally contain 12% aluminum, 16% aluminum, and 16% aluminum with 3.5% molybdenum, respectively. These three alloys are classified as magnetically soft materials; that is, they become magnetized in a magnetic field but are easily demagnetized when the field is removed.

The addition of more than 8% aluminum to iron results in alloys that are too brittle for many uses because of difficulties in fabrication. However, addition of aluminum to iron markedly increases its resistance to oxidation. One steel containing 6% aluminum possesses good oxidation resistance up to 2300°F (1300°C). *See* ALUMINUM ALLOYS.

Iron-carbon alloys. The principal iron-carbon alloys are wrought iron, cast iron, and steel.

Wrought iron of good quality is nearly pure iron; its carbon content seldom exceeds 0.035%. In addition, it contains 0.075–0.15% silicon, 0.10 to less than 0.25% phosphorus, less than 0.02% sulfur, and 0.06–0.10% manganese. Not all of these elements are alloyed with the iron; part of them may be associated with the intermingled slag that is a characteristic of this product. Because of its low carbon content, the properties of wrought iron cannot be altered in any useful way by heat treatment. *See* WROUGHT IRON.

Cast iron may contain 2–4% carbon and varying amounts of silicon, manganese, phosphorus, and sulfur to obtain a wide range of physical and mechanical properties. Alloying elements (silicon, nickel, chromium, molybdenum, copper, titanium, and so on) may be added in amounts varying from a few tenths to 30% or more. Many of the alloyed cast irons have proprietary compositions. *See* CAST IRON; CORROSION.

Steel is a generic name for a large group of iron alloys that include the plain carbon and alloy steels. The plain carbon steels represent the most important group of engineering materials known. Although any iron-carbon alloy containing less than about 2% carbon can be considered a steel, the American Iron and Steel Institute (AISA) standard carbon steels embrace a range of carbon contents from 0.06% maximum to about 1%. In the early days of the American steel industry, hundreds of steels with different chemical compositions were produced to meet individual demands of purchasers. Many of these steels differed only slightly from each other in chemical composition. Studies were undertaken to provide a simplified list of fewer steels that would still serve the varied needs of fabricators and users of steel products. The Society of Automotive Engineers (SAE) and the AISI both were prominent in this effort, and both periodically publish lists of steels, called standard steels, classified by chemical composition. These lists are published in the *SAE Handbook* and the AISI's *Steel Products Manuals*. The lists are altered periodically to accommodate new steels and to provide for changes in consumer requirements. There are minor differences between some of the steels listed by the AISI and SAE. Only the AISI lists will be considered here. The standard steels represent a large percentage of all steel produced and, although considerably fewer in number, have successfully replaced the large number of specialized compositions formerly used.

A numerical system is used to indicate grades of standard steels. Provision also is made to use certain letters of the alphabet to indicate the steel-making

TABLE 2. AISI standard steel designations

Type	Series designation*	Composition
Carbon steels	10xx	Nonresulfurized carbon steel
	11xx	Resulfurized carbon steel
	12xx	Rephosphorized and resulfurized carbon steel
Constructional alloy steels	13xx	Manganese 1.75%
	23xx	Nickel 3.50%
	25xx	Nickel 5.00%
	31xx	Nickel 1.25%, chromium 0.65%
	33xx	Nickel 3.50%, chromium 1.55%
	40xx	Molybdenum 0.25%
	41xx	Chromium 0.50 or 0.95%, molybdenum 0.12 or 0.20%
	43xx	Nickel 1.80%, chromium 0.50 or 0.80%, molybdenum 0.25%
	46xx	Nickel 1.55 or 1.80%, molybdenum 0.20 or 0.25%
	47xx	Nickel 1.05%, chromium 0.45%, molybdenum 0.20%
	48xx	Nickel 3.50%, molybdenum 0.25%
	50xx	Chromium 0.28 or 0.40%
	51xx	Chromium 0.80, 0.90, 0.95, 1.00 or 1.05%
	5xxxx	Carbon 1.00%, chromium 0.50, 1.00, or 1.45%
	61xx	Chromium 0.80 or 0.95%, vanadium 0.10 or 0.15% min
	86xx	Nickel 0.55%, chromium 0.50 or 0.65%, molybdenum 0.20%
	87xx	Nickel 0.55%, chromium 0.50%, molybdenum 0.25%
	92xx	Manganese 0.85%, silicon 2.00%
	93xx	Nickel 3.25%, chromium 1.20%, molybdenum 0.12%
	98xx	Nickel 1.00%, chromium 0.80%, molybdenum 0.25%
Stainless and heat-resisting steels	2xx	Chromium-nickel-manganese steels; nonhardenable, austenitic, and nonmagnetic
	3xx	Chromium-nickel steels; nonhardenable, austenitic, and nonmagnetic
	4xx	Chromium steels of two classes: one class hardenable, martensitic, and magnetic; the other nonhardenable, ferritic, and magnetic
	5xx	Chromium steels; low chromium heat resisting

*The x's are replaced by actual numerals in defining a steel grade, as explained in the text.

process, certain special additions, and steels that are tentatively standard, but these are not pertinent to this discussion. **Table 2** gives the basic numerals for the AISI classification and the corresponding types of steels. In this system the first digit of the series designation indicates the type to which a steel belongs; thus 1 indicates a carbon steel, 2 indicates a nickel steel, and 3 indicates a nickel-chromium steel. In the case of simple alloy steels, the second numeral usually indicates the percentage of the predominating alloying element. Usually, the last two (or three) digits indicate the average carbon content in points, or hundredths of a percent. Thus, 2340 indicates a nickel steel containing about 3% nickel and 0.40% carbon.

All carbon steels contain minor amounts of manganese, silicon, sulfur, phosphorus, and sometimes other elements. At all carbon levels the mechanical properties of carbon steel can be varied to a useful degree by heat treatments that alter its microstructure. Above about 0.25% carbon steel can be hardened by heat treatment. However, most of the carbon steel produced is used without a final heat treatment. *See* HEAT TREATMENT (METALLURGY).

Alloy steels are steels with enhanced properties attributable to the presence of one or more special elements or of larger proportions of manganese or silicon than are present ordinarily in carbon steel. The major classifications of alloy steels are high-strength, low-alloy; AISI alloy; alloy tool; heat-resisting; electrical; and austenitic manganese. Some of these iron alloys are discussed briefly in this article; for more detailed attention *see* STAINLESS STEEL; STEEL.

Iron-chromium alloys. An important class of iron-chromium alloys is exemplified by the wrought stainless and heat-resisting steels of the type 400 series of the AISI standard steels, all of which contain at least 12% chromium, which is about the minimum chromium content that will confer stainlessness. However, considerably less than 12% chromium will improve the oxidation resistance of steel for service up to 1200°F (650°C), as is true of AISI types 501 and 502 steels that nominally contain about 5% chromium and 0.5% molybdenum. A comparable group of heat- and corrosion-resistant alloys, generally similar to the 400 series of the AISI steels, is covered by the Alloy Casting Institute specifications for cast steels.

Corrosion-resistant cast irons alloyed with chromium contain 12–35% of that element and up to 5% nickel. Cast irons classified as heat-resistant contain 15–35% chromium and up to 5% nickel.

During World War I a high-carbon steel used for making permanent magnets contained 1–6% chromium (usually around 3.5%); it was developed to replace the magnet steels containing tungsten that had been formerly but could not then be made because of a shortage of tungsten. *See* MAGNET.

Iron-chromium-nickel alloys. The wrought stainless and heat-resisting steels represented by the type 200 and the type 300 series of the AISI standard steels are an important class of iron-chromium-nickel alloys. A comparable series of heat- and corrosion-resistant alloys is covered by specifications of the Alloy Casting Institute. Heat- and corrosion-resistant cast irons contain 15–35% chromium and up to 5% nickel.

Iron-chromium-aluminum alloys. Electrical-resistance heating elements are made of several iron alloys of this type. Nominal compositions are 72% iron, 23% chromium, 5% aluminum; and 55% iron, 37.5% chromium, 7.5% aluminum. The iron-chromium-aluminum alloys (with or without 0.5–2% cobalt) have higher electrical resistivity and lower density than nickel-chromium alloys used for the same purpose. When used as heating elements in furnaces, the iron-chromium-aluminum alloys can be operated at temperatures of 4262°F (2350°C) maximum. These alloys are somewhat brittle after elevated temperature use and have a tendency to grow or increase in length while at temperature, so that heating elements made from them should have additional mechanical support. Addition of niobium (columbium) reduces the tendency to grow.

Because of its high electrical resistance, the 72% iron, 23% chromium, 5% aluminum alloy (with 0.5% cobalt) can be used for semiprecision resistors in, for example, potentiometers and rheostats. *See* ELECTRICAL RESISTANCE; RESISTANCE HEATING.

Iron-cobalt alloys. Magnetically soft iron alloys containing up to 65% cobalt have higher saturation values than pure iron. The cost of cobalt limits the use of these alloys to some extent. The alloys also are characterized by low electrical resistivity and high hysteresis loss. Alloys containing more than 30% cobalt are brittle unless modified by additional alloying and special processing. Two commercial alloys with high permeability at high field strengths (in the annealed condition) contain 49% cobalt with 2% vanadium, and 35% cobalt with 1% chromium. The latter alloy can be cold-rolled to a strip that is sufficiently ductile to permit punching and shearing. In the annealed state, these alloys can be used in either ac or dc applications. The alloy of 49% cobalt with 2% vanadium has been used in pole tips, magnet yokes, telephone diaphragms, special transformers, and ultrasonic equipment. The alloy of 35% cobalt with 1% chromium has been used in high-flux-density motors and transformers as well as in some of the applications listed for the higher-cobalt alloy.

Although seldom used now, two high-carbon alloys called cobalt steel were used formerly for making permanent magnets. These were both high-carbon steels. One contained 17% cobalt, 2.5% chromium, 8.25% tungsten; the other contained 36% cobalt, 5.75% chromium, 3.75% tungsten. These are considered magnetically hard materials as compared to the magnetically soft materials discussed in this article.

Iron-manganese alloys. The important commercial alloy in this class is an austenitic manganese steel (sometimes called Hadfield manganese steel after its inventor) that nominally contains 1.2% carbon and 12–13% manganese. This steel is highly resistant to abrasion, impact, and shock.

Iron-nickel alloys. The iron-nickel alloys discussed here exhibit a wide range of properties related to their nickel contents.

Nickel content of a group of magnetically soft materials ranges from 40 to 60%; however, the highest saturation value is obtained at about 50%. Alloys with nickel content of 45–60% are characterized by high permeability and low magnetic losses. They are used in such applications as audio transformers, magnetic amplifiers, magnetic shields, coils, relays, contact rectifiers, and choke coils. The properties of the alloys can be altered to meet specific requirements by special processing techniques involving annealing in hydrogen to minimize the effects of impurities, grain-orientation treatments, and so on.

Another group of iron-nickel alloys, those containing about 30% nickel, is used for compensating changes that occur in magnetic circuits due to temperature changes. The permeability of the alloys decreases predictably with increasing temperature.

Low-expansion alloys are so called because they have low thermal coefficients of linear expansion. Consequently, they are valuable for use as standards of length, surveyors' rods and tapes, compensating pendulums, balance wheels in timepieces, glass-to-metal seals, thermostats, jet-engine parts, electronic devices, and similar applications.

The first alloy of this type contained 36% nickel with small amounts of carbon, silicon, and manganese (totaling less than 1%). Subsequently, a 39% nickel alloy with a coefficient of expansion equal to that of low-expansion glasses and a 46% nickel alloy with a coefficient equal to that of platinum were developed. Another important alloy is one containing 42% nickel that can be used to replace platinum as lead-in wire in light bulbs and vacuum tubes by first coating the alloy with copper. An alloy containing 36% nickel and 12% chromium has a constant modulus of elasticity and low expansivity over a broad range of temperatures. Substitution of 5% cobalt for 5% nickel in the 36% nickel alloy decreases its expansivity. Small amounts of other elements affect the coefficient of linear expansion, as do variations in heat treatment, cold-working, and other processing procedures.

A 9% nickel steel is useful in cryogenic and similar applications because of good mechanical properties at low temperatures. Two steels (one containing 10–12% nickel, 3–5% chromium, about 3% molybdenum, and lesser amounts of titanium and aluminum and another with 17–19% nickel, 8–9% cobalt, 3–3.5% molybdenum, and small amounts of titanium and aluminum) have exceptional strength in the heat-treated (aged) condition. These are known as maraging steels.

Cast irons containing 14–30% nickel and 1.75–5.5% chromium possess good resistance to heat and corrosion. *See* THERMAL EXPANSION.

Iron-silicon alloys. There are two types of iron-silicon alloys that are commercially important: the magnetically soft materials designated silicon or electrical steel, and the corrosion-resistant, high-silicon cast irons.

Most silicon steels used in magnetic circuits contain 0.5–5% silicon. Alloys with these amounts of silicon have high permeability, high electrical resistance, and low hysteresis loss compared with relatively pure iron. Most silicon steel is produced in flat-rolled (sheet) form and is used in transformer cores, stators, and rotors of motors, and so on that are built up in laminated-sheet form to reduce eddy-current losses. Silicon-steel electrical sheets, as they are called commercially, are made in two general classifications: grain-oriented and nonoriented.

The grain-oriented steels are rolled and heat-treated in special ways to cause the edges of most of the unit cubes of the metal lattice to align themselves in the preferred direction of optimum magnetic properties. Magnetic cores are designed with the main flux path in the preferred direction, thereby taking advantage of the directional properties. The grain-oriented steels contain about 3.25% silicon, and they are used in the highest-efficiency distribution and power transformers and in large turbine generators.

The nonoriented steels may be subdivided into low-, intermediate-, and high-silicon classes. Low-silicon steels contain about 0.5–1.5% silicon and are used principally in rotors and stators of motors and generators; steels containing about 1% silicon are also used for reactors, relays, and small intermittent-duty transformers. Intermediate-silicon steels contain about 2.5–3.5% silicon and are used in motors and generators of average to high efficiency and in small- to medium-size intermittent-duty transformers, reactors, and motors. High-silicon steels contain about 3.75–5% silicon and are used in power transformers and communications equipment in highest efficiency motors, generators, and transformers.

High-silicon cast irons containing 14–17% silicon and sometimes up to 3.5% molybdenum possess corrosion resistance that makes them useful for acid-handling equipment and for laboratory drain pipes.

Iron-tungsten alloys. Although tungsten is used in several types of relatively complex alloy (including high-speed steels not discussed here), the only commercial alloy made up principally of iron and tungsten was a tungsten steel containing 0.5% chromium in addition to 6% tungsten that was used up to the time of World War I for making permanent magnets.

Hard-facing alloys. Hard-facing consists of welding a layer of metal of special composition on a metal surface to impart some special property not possessed by the original surface. The deposited metal may be more resistant to abrasion, corrosion, heat, or erosion than the metal to which it is applied. A considerable number of hard-facing alloys are available commercially. Many of these would not be considered iron alloys by the 50% iron content criterion adopted for the iron alloys in this article, and they will not be discussed here. Among the iron alloys are low-alloy facing materials containing chromium as the chief alloying element, with smaller amounts of manganese, silicon, molybdenum, vanadium, tungsten, and in some cases nickel to make a total alloy content of up to 12%, with the balance iron. High-alloy ferrous materials containing a total of 12–25% alloying elements form another group of hard-facing alloys; a third group contains 26–50% alloying elements. Chromium, molybdenum, and manganese are the principal alloying elements in the 12–25% group; smaller amounts of molybdenum, vanadium, nickel, and in some cases titanium are present in various proportions. In the 26–50% alloys, chromium (and in some cases tungsten) is the principal alloying element, with manganese, silicon, nickel, molybdenum, vanadium, niobium (columbium), and boron as the elements from which a selection is made to bring the total alloy content within the 26–50% range.

Permanent-magnet alloys. These are magnetically hard ferrous alloys, many of which are too complex to fit the simple compositional classification used above for other iron alloys. As already mentioned in discussing iron-cobalt and iron-tungsten alloys, the high-carbon steels (with or without alloying elements) are now little used for permanent magnets. These have been supplanted by a group of sometimes complex alloys with much higher retentivities. The ones considered here are all proprietary compositions. Two of the alloys contain 12% cobalt and 17% molybdenum and 12% cobalt and 20% molybdenum.

Members of a group of six related alloys contain iron, nickel, aluminum, and with one exception cobalt; in addition, three of the cobalt-containing alloys contain copper and one has copper and titanium. Unlike magnet steels, these alloys resist demagnetization by shock, vibration, or temperature variations. They are used in magnets for speakers, watt-hour meters, magnetrons, torque motors, panel and switchboard instruments, and so on, where constancy and degree of magnet strength are important. *See* ALLOY STRUCTURES; MAGNETISM.

Harold E. McGannon

Bibliography. American Society for Testing and Materials, *Annual Book of ASTM Standards*, 1992, vol. 01.02: *Ferrous Castings; Ferroalloys*, 1992; F. E. Fletcher, *High-Strength Low-Alloy Steels: Status*, *Selection*, *and Physical Metallurgy*, 1982; R. A. Oriani et al. (eds.), *Hydrogen Dehydration of Ferrous Alloys*, 1985; Y. S. Touloukian and C. Y. Ho (eds.), *Properties of Selected Ferrous Alloying Elements*, vol. 3, 1989.

Iron metabolism

A nearly continuous cycle whereby iron from organic and inorganic sources is converted to form iron-porphyrin compounds, which can be utilized by the body. One such compound, termed a heme, is hemoglobin; more than 60% of the iron in the body is used in hemoglobin metabolism. Iron is also essential for other heme compounds, such as myoglobin and cytochromes, and for a wide variety of nonheme enzymes, including many in the citric acid cycle.

Hemoglobin metabolism involves formation and breakdown. Formation is initiated by the process of absorption. Unlike most elements, iron has no specific mechanism for excretion so that absorption must be closely monitored to control body content and ensure replacement of the daily loss. The normal daily requirements of iron are 1 mg for males, 2 mg for young women (to cover the loss from menstruation), and 3 mg for pregnant women. From an average diet, an absorption rate of 5–10% is possible. Absorption occurs best from animal and plant hemes and less from inorganic ferric salts; different mechanisms are involved for the two types. Many factors influence absorption, especially gastric acid, which can solubilize iron salts and prevent their precipitation in the duodenum.

In the duodenum, iron rapidly enters the mucosal cells of the intestinal villi, where the iron is released from heme by the enzyme heme oxygenase. There, the ferrous iron destined for the formation of hemoglobin in the developing red cells of the marrow (the erythroblasts) is converted to ferric ions by ceruloplasmin and is attached to the transport glycoprotein transferrin. The nonassimilated iron remains in the intestinal cell and is combined with the storage protein apoferritin to form ferritin, which is lost by the body when the mucosal cells are shed after their 3–5-day life cycle. The mechanism by which the mucosal cell knows what to discard and what to assimilate is unknown. Absorption is increased by iron deficiency, anemia of many varieties, hypoxia, and increased red cell production (erythropoiesis).

Once ferric iron is attached to transferrin, it circulates in the blood until it attaches to transferrin receptors on immature red blood cells in the marrow. There may be as many as 300,000 receptors on the surface of an individual cell. Once attached, the transferrin receptor complex is taken in by endocytosis, and the iron is released and reduced to ferrous ions. The transferrin receptor complex returns to the cell surface, and the transferrin reenters the blood plasma. The iron enters the mitochondria and is inserted into protoporphyrin by the enzyme ferrochetalase to form heme, which when combined with the protein globin forms the respiratory pigment hemoglobin.

Mature red blood cells cannot take up iron; at the end of their 120-day life span they are engulfed by the monocyte-macrophage cells in liver and spleen, where the iron is released by the enzyme heme oxygenase. Sixty percent of this iron is rapidly returned to the marrow to produce red blood cells, while the remainder is stored as ferritin in the labile pool for release as needed. Iron excessively absorbed from the gut or released from the labile pool and not destined for the formation of red blood cells may enter the storage compartment as ferritin or hemosiderin. Apoferritin, the iron-free storage protein, exists in most living cells, and ferric ions can be stored in its hollow sphere to form a growing crystal of ferric oxyhydroxide (FeOOH). Small amounts of apoferritin enter the circulatory system at levels directly paralleling those of the stores. Hemosiderin occurs in the monocyte macrophages of the liver and spleen as FeOOH stripped of its apoferritin shell. Both storage forms may be rapidly released on demand for increased red blood cell production. Excessive amounts of iron may be stored as hemosiderin following multiple transfusions of blood in non-iron-deficient anemias, in the disease hemochromatosis, and following undue breakdown of cells. *See* BLOOD; HEMOGLOBIN.

John Murray

Bibliography. W. J. Williams et al. (eds.), *Hematology*, 5th ed., 1995.

Iron metallurgy

Extracting iron from ores and preparing it for use. Extraction involves the conversion of naturally occurring iron-bearing minerals into metallic iron. The term ironmaking is commonly used to include all of the industrial processes that convert raw materials into iron. The major process for the production of iron is the iron blast furnace, which varies widely in size and specific features. However, the growth of alternative direct-reduction processes has been very significant (**Fig. 1**). The principal difference between the blast furnace process and the direct-reduction processes is the temperature of operation. In the blast furnace, high operating temperatures enable the production of molten iron. At the lower operating temperatures of the direct-reduction processes, solid or sponge iron is produced. Most of the iron produced in the world is used in the production of steel. The remainder is converted to iron castings, ferroalloys, and iron powder.

An alternative method for ironmaking is direct smelting. These processes rely on the direct reaction of noncoking coal, oxygen, and iron ore in a high-temperature smelter to produce molten iron. In this sense, they fall between the blast furnace and direct reduction (Fig. 1). The primary impetus for these processes is the elimination of coking in the production of iron from coal. *See* FERROALLOY.

Ore preparation. The major ore deposits contain iron in the form of oxides, such as hematite (Fe_2O_3) and magnetite (Fe_3O_4). These oxides are physically separated from the other constituents in the ore by sequences of operations, including crushing and fine grinding, magnetic and flotation concentration, and dewatering. The separation of iron oxide concentrates in this manner is a branch of a technology called mineral beneficiation or ore dressing. The iron oxide particles in the concentrates are almost as fine in size as talcum powder (−200 mesh). The gangue constituents of the ore are also finely divided, and this has caused major problems in the development of environmentally suitable disposal methods. The iron oxide concentrates must be consolidated into hard pellets, sintered lumps, or briquettes to become suitable charge materials for the iron blast furnace and most of the direct-reduction processes. In general, the ore dressing and consolidation operations

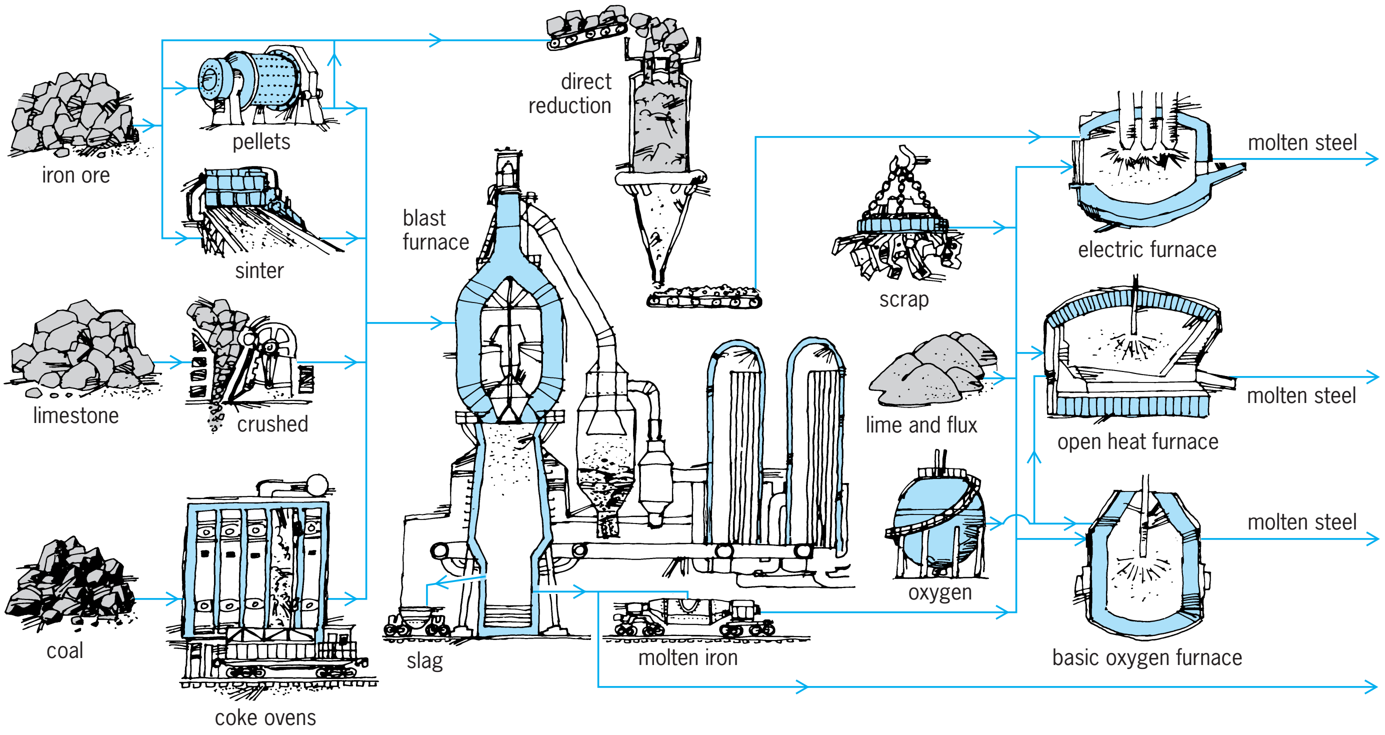


Fig. 1. **The position of ironmaking processes (blast furnace and direct reduction) in the steelmaking sequence. (*American Iron and Steel Institute*)**

are performed near the mining sites. *See* FLOTATION; ORE DRESSING.

Reduction of oxide concentrates. The conversion of iron oxide to metallic iron is accomplished by the application of heat and reducing agents. In sequence, hematite (Fe_2O_3), magnetite (Fe_3O_4), and wüstite (FeO) are reduced by carbon monoxide (CO) and hydrogen (H_2) as shown in reactions (1)–(3).

$$\tfrac{1}{2}Fe_2O_3 + \tfrac{1}{6}CO \rightarrow \tfrac{1}{3}Fe_3O_4 + \tfrac{1}{6}CO_2 \qquad (1a)$$

$$\tfrac{1}{2}Fe_2O_3 + \tfrac{1}{6}H_2 \rightarrow \tfrac{1}{3}Fe_3O_4 + \tfrac{1}{6}H_2O \qquad (1b)$$

$$\tfrac{1}{3}Fe_3O_4 + \tfrac{1}{3}CO \rightarrow FeO + \tfrac{1}{3}CO_2 \qquad (2a)$$

$$\tfrac{1}{3}Fe_3O_4 + \tfrac{1}{3}H_2 \rightarrow FeO + \tfrac{1}{3}H_2O \qquad (2b)$$

$$FeO + CO \rightarrow Fe + CO_2 \qquad (3a)$$

$$FeO + H_2 \rightarrow Fe + H_2O \qquad (3b)$$

The sum of reactions (1*a*), (2*a*), (3*a*) yields reaction (4), which represents an overall reaction for the

$$\tfrac{1}{2}Fe_2O_3 + \tfrac{3}{2}CO \rightarrow Fe + \tfrac{3}{2}CO_2 \qquad (4)$$

reduction of hematite to metallic iron with carbon monoxide.

The reducing agents, CO and H_2, are derived from coal, natural gas, oil, wood, and other carbonaceous or organic materials. The manner in which the reducing agents are obtained from these raw materials is complex. For example, the iron blast furnace process uses coke (Fig. 1). When coal is heated in coke ovens, a fuel gas is released, leaving a residue of hard carbonaceous material called coke. The released gas is primarily methane (CH_4) with small amounts of other recoverable organic species. After processing, the residual gas is a useful fuel. The hard coke charged at the top of the blast furnace descends slowly to the bosh region (**Fig. 2**), where it is burned with hot air blasted through the tuyeres. In the presence of excess coke the combustion reaction is given by reaction (5). In most blast furnaces, hydrocarbons

$$C + \tfrac{1}{2}O_2 \rightarrow CO \qquad (5)$$

(oil, gas, tar, and so on) are added to the hot blast to provide H_2, which reduces the required amount of coke to produce a unit of iron. For example, where methane is added to the hot blast, hydrogen is generated by reaction (6). In addition, small but significant

$$CH_4 + \tfrac{1}{2}O_2 \rightarrow CO + 2H_2 \qquad (6)$$

amounts of CO and H_2 are regenerated from CO_2 and H_2O at higher levels in the stack by solution loss reactions represented by reactions (7).

$$CO_2 + C \rightarrow 2CO \qquad (7a)$$

$$H_2O + C \rightarrow CO + H_2 \qquad (7b)$$

The sources of the CO_2 and H_2O for these reactions are reactions (1)– (3), which occur simultaneously. As the gases and solids move countercurrently, the descending solid oxides release their oxygen to the ascending gas stream by means of the reduction reactions. The solution loss reactions provide some regeneration of CO and H_2 to maintain the reducing strength of the ascending gases. *See* COKE.

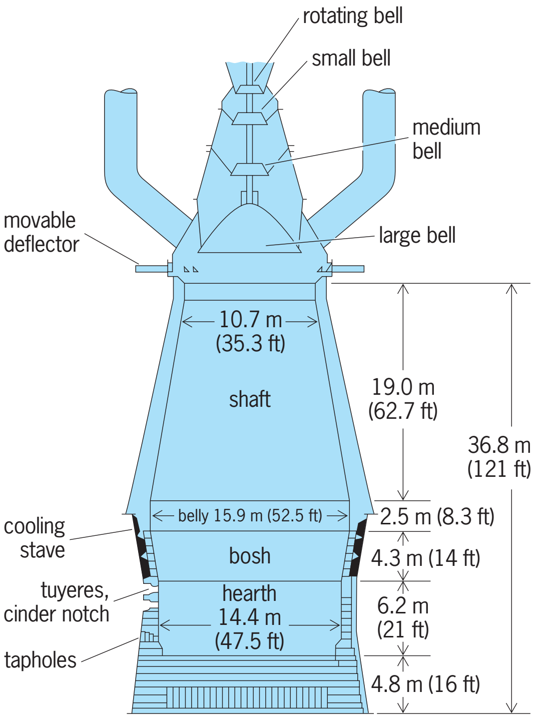


Fig. 2. Iron blast furnace. (*After J. G. Peacey and W. G. Davenport, The Iron Blast Furnace, Pergamon, 1979*)

For the direct-reduction processes, CO and H_2 are produced from the available raw materials in a variety of ways. Several of the rotary kiln processes utilize the direct charging of coal with iron oxide concentrates. The fixed carbon and the volatile matter of the coal are partially burned in place with the iron oxides to generate the reducing gases. Several of the shaft, retort, and fluidized-bed processes utilize the separate generation of reducing gas from natural gas.

In some direct smelting processes, pulverized coal is injected into a molten iron/slag emulsion to provide reductant (fixed carbon and hydrocarbon gases). Simultaneously, iron oxide fines and pellets are fed into the slag, where they dissolve readily. An oxygen lance provides oxidant to burn hydrocarbons and fixed carbon. The overall reaction system in direct smelting is basically the same as that in the blast furnace; however, it is as if the blast furnace were compressed into just the bosh and hearth regions. The process depends upon the fluid dynamic conditions in the iron/slag emulsion, also referred to as a foam.

Material balances. The principal raw materials used in the extraction of iron are oxide concentrates, reducing agents, air, water and steam, auxiliary fuels used directly or converted to electrical energy, and fluxes (**Fig. 3**). The most common flux used in the iron blast furnace is limestone ($CaCO_3$). Large modern furnaces are extremely efficient and values of coke consumption as low as 0.40 kg of coke per kilogram of iron (800 lb/short ton) have been achieved with oil injection. The material balance provides an indication of the relative magnitudes of the major input and output items. Direct-reduction and direct-smelting processes have features in common with the blast furnace; however, the input and output statements may appear to be different to an experienced observer.

The quantity of air required per unit of iron produced in the blast furnace is an impressive figure. At least 1 m^3 of air must be blown into the blast furnace or direct-reduction process for each kilogram of iron produced. The exhaust gases from the blast furnace contain considerable amounts of H_2 and CO.

In modern plants, material balances are monitored by sophisticated computer techniques. It is essential to keep the system operating productively and efficiently. In an integrated steel plant, interruptions in the production of iron from the blast furnace can upset the entire sequence of operations. In this respect, the integrated steel plant is similar to the assembly line for any manufactured product. *See* PRESSURIZED BLAST FURNACE.

Energy balances and temperature profiles. Total energy consumption in steelmaking is much lower than in the production of other common metals (**Table 1**). To some extent the higher thermal efficiency of ironmaking may be attributed to the large scale of iron and steel operations. Another important factor is that recycling and reuse of waste materials and energy sources have been extensively developed (**Table 2**).

Estimated temperature profiles have been established (**Fig. 4**). The temperature of the descending solids rises rapidly to about 1000 K (1340°F) in the top portion of the furnace, rises gradually

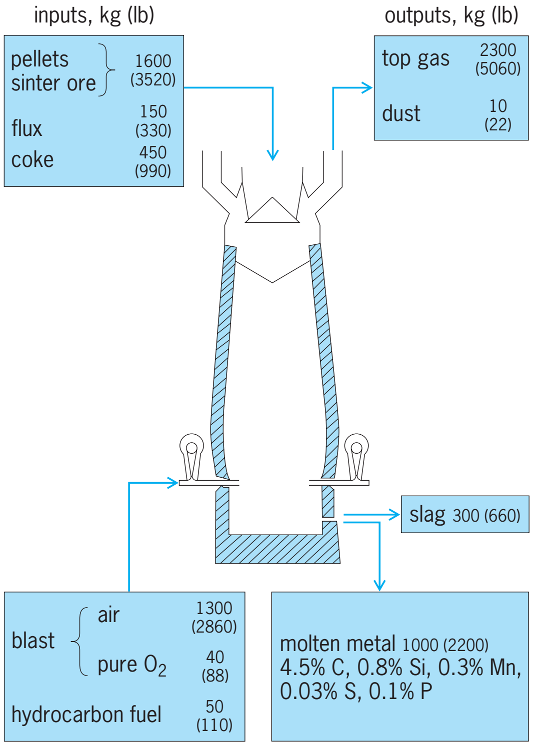


Fig. 3. Simplified material balance for an iron blast furnace. (*After J. G. Peacey and W. G. Davenport, The Iron Blast Furnace, Pergamon, 1979*)

TABLE 1. Approximate energy requirements for the extraction of metals from ores

Metal	Theoretical, 10^2 kJ/kg	Actual, 10^3 kJ/kg
Iron	3.7	19.4
Copper	1.6	54.0
Aluminum	18.3	236.0
Titanium	12.0	501.0

TABLE 2. Approximate simplified energy balance for a blast furnace

Parameter	10^3 kJ/kg
Energy input	
Calorific value of coke	14.87
Sensible heat of hot air blast	1.95
Calorific value of injected fuel	1.12
Total	17.94
Energy output	
Calorific value of recoverable top gas	7.40
Energy consumed	
Heat for reactions, sensible heat of slag and iron, and losses	10.54

to about 1200 K (1700°F) in the central portion, and rises rapidly to temperatures in excess of the fusion point in the lower portion. The radial temperature gradient is very steep at the midheight of the furnace, where the temperature at the vertical axis may be as high as 1900 K (2960°F), while the wall temperature is only 1200 K (1700°F). The residence time of the ascending gases in the furnace is of the order of 1 s, while that of the descending solids is of the order of 10^4 s. Uniformity of gas and solid movement in the furnace is highly desirable. Excessive channeling, erratic flow, and formation of static zones can lead to major losses in efficiency and productivity.

Composition of products. Although meteoric iron containing nickel was used before 4000 B.C., it is believed that iron was not extracted from ore until about 3000 B.C. The earliest extracted form and composition of iron was similar to that produced in modern direct-reduction processes that operate at temperatures below the melting point of iron. This reduced solid iron is called sponge iron because of its porosity. Sponge iron contains significant amounts of trapped carbon and unreduced oxide, and it must be worked severely or melted and refined. Blast furnaces produce liquid iron that is commonly called hot metal or pig iron. Pig iron contains more than 4% carbon and significant amounts of manganese, silicon, and phosphorus. Pig iron must be refined in subsequent steelmaking operations.

Trends. The largest modern furnaces can produce more than 2.2×10^7 lb (10^7 kg) of iron per day, operating with careful charge control, high top pressure, high blast temperature, and fuel injection. It has become common practice to lower the quantity of slag and its basicity. Sulfur control is provided through the addition of a separate process for the desulfurization of hot metal. With the development of advanced sensors and mathematical modeling methods and supercomputers, it has become possible to predict and control the operation of large blast furnaces to an extent far greater than was previously the case. In seconds, analyses of variations can be made, and corrective actions can be taken to maintain high productivity and efficiency. *See* SIMULATION; SUPERCOMPUTER.

The modern iron blast furnace is a highly developed, very efficient, and economical process for large steel plants. However, it requires coke, which must be produced in associated coke ovens, a practice associated with environmental problems; and it is very large and inflexible. Once it becomes operational, the blast furnace goes continuously, and the associated steel plant activities must be matched to take its output. Further, the blast furnace offers little opportunity for incremental increases in steel plant capacity. A major expansion and investment is required for installation of a new blast furnace. A typical large modern integrated steel plant operates with one to four blast furnaces.

Direct-reduction processes have been constructed at many sites throughout the world, particularly in the developing countries. The direct reduction processes offer the alternative of a shippable iron product suitable for steel production in small- and large-scale steel plants. In effect, the direct-reduction iron products act as a substitute for scrap (Fig. 1) with the added advantage that they are cleaner and less contaminated with tramp elements than bundles of commercial scrap. The daily production rates of the direct-reduction processes are less than 2.2×10^6 lb

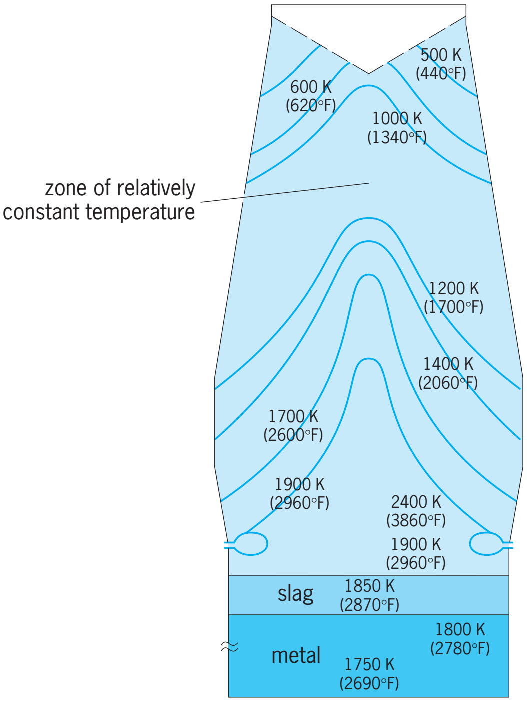


Fig. 4. Temperature profiles in a modern blast furnace. (*After J. G. Peacey and W. G. Davenport, The Iron Blast Furnace, Pergamon, 1979*)

(1×10^6 kg) and often are as low as 4.2×10^5 lb (2×10^5 kg). The direct smelting processes are designed to be flexible and compact with high production intensity (tons of iron produced per unit of working reactor volume) and to start up quickly and adjust production rate to demand. They are adaptable to small-scale steel production facilities. However, molten iron is not shippable and must be transferred to nearby steelmaking furnaces or cast into pigs. *See* CAST IRON; IRON; IRON ALLOYS; PYROMETALLURGY; STEEL MANUFACTURE. George R. St. Pierre

Bibliography. American Institute of Mining, Metallurgical, and Petroleum Engineers (AIME), *Blast Furnace*: *Theory and Practice*, 1969; AIME, *Ironmaking Proceedings*, annually; Gmelin Institute for Inorganic Chemistry of the Max Planck Society for the Advancement of Science, *Metallurgy of Iron*, 4th ed., vols. 1–12, 1992; R. J. Fruehan et al. (eds.), *Making, Shaping, and Treating of Steel*, 11th ed., 1998; W.-K. Lu, A critical review of blast furnace reactions, *Iron Steelmak.*, pp. 51–62, October 1987; T. P. McAloon, Alternate ironmaking update, *Iron Steelmak.*, pp. 37–56, February 1994; J. G. Peacey and W. G. Davenport, *The Iron Blast Furnace*, 1979.

Ironwood

The name given to any of at least 10 kinds of tree in the United States. Because of uncertainty as to just which tree is indicated, the name ironwood has been abandoned in the checklist of the native and naturalized trees of the United States. Probably the best known of the 10 is the American hornbeam (*Carpinus caroliniana*). Some of the others are *Ostrya virginiana*, eastern hophornbeam; *Bumelia lycioides*, buckthorn bumelia; *B. tenax*, tough bumelia; *Cliftonia monophylla*, buckwheat tree; and *Cyrilla racemiflora*, swamp cyrilla or swamp ironwood. All of these species except *Ostrya* are restricted to the southeastern United States. Others commonly called ironwood are *Eugenia confusa*, redberry eugenia of southern Florida and the Florida Keys; *Exothea paniculata*, butterbough of southern Florida; and *Ostrya knowltonii*, Knowlton hophornbeam of southwestern United States. Leadwood (*Krugiodendron ferreum*), a native of southern Florida, has the highest specific gravity of all woods native to the United States and is also known as black ironwood. *See* FOREST AND FORESTRY; HOPHORNBEAM; HORNBEAM; TREE.

Arthur H. Graves; Kenneth P. Davis

Irregularia

The name given by P. A. Latreille in 1825 to what is now recognized as a natural (monophyletic, that is, evolved from a common ancestral form) group of echinoids (sea urchins). Previously, workers such as J. Durham and R. Melville suggested that the group was polyphyletic (evolved from different ancestral lineages), a proposal followed by other influential works. However, this suggestion has been rejected by more recent workers such as M. Jensen, A. Smith, and T. Saucède and colleagues who showed in phylogenetic analyses that all irregular echinoids are derived from eodiadematid-like forms in the Lower Jurassic. Most Irregularia, and all the extant forms, are exocyclic: the periproct (the area surrounding the anus) plus the anus is displaced in the direction of interambulacrum 5 (Lovén's numbering system) to become situated at the edge of or outside the apical system. (The interambulacrum is the area between two ambulacra, which are the radial series of plates along which the tube feet are arranged.) Therefore, these Irregularia exhibit an anterior-posterior axis running down the centers of ambulacrum III and interambulacrum 5, inducing secondary bilateral symmetry. Bilateral symmetry (symmetry along a central axis, with division into equivalent left and right halves) is most pronounced in the spatangoid heart urchins, and in stark contrast to the radial symmetry (symmetry about a central point) of the "regular" urchins. Additional characteristics of the Irregularia include the miniaturization of external appendages [spines, tube feet, pedicellariae (small grasping organs)], the relatively large size of the periproct, the high density of primary tubercles, and in most forms a relatively low test (the internal, limestone skeleton of sea urchins) profile. The Irregularia includes a basal group, the Eodiadematidae, plus holectypoids, pygasteroids, cassiduloid and clypeasteroid neognathostomates, and holasteroid and spatangoid atelostomates. Irregular echinoids left a richer fossil record than the regular urchins, in part because the test of irregulars is less prone to postmortem disintegration and because they tended to live in softer substrata that led to better preservation. Irregular echinoids are largely sediment-swallowing "podial particle pickers" and have developed sophisticated mechanisms for handling fine sediments. *See* ECHINOIDA; ECHINOIDEA; ECHINODERMATA; HOLECTYPOIDA; NEOGNATHOSTOMATA; PYGASTEROIDA. Rich Mooi

Bibliography. J. W. Durham, Phylogeny and evolution, in R. C. Moore (ed.), Part U: *Treatise on Invertebrate Paleontology, Echinodermata-Echinozoa-Echinoidea*, vol. 3, pp. 266–269, 1966; J. W. Durham and R. V. Melville, A classification of echinoids, *J. Paleontol.*, 31:242–272, 1957; M. Jensen, Morphology and classification of the Euechinoidea Bronn, 1860—a cladistic analysis, *Videnskabelige Meddelelser fra Dansk Naturhistorisk Forening i Kjobenhavn*, 143:7–99, 1981; A. B. Smith, *Echinoid Palaeobiology*, Allen & Unwin, London, 1984.

Irrigation (agriculture)

The artificial application of water to the soil to produce plant growth. Irrigation also cools the soil and atmosphere, making the environment favorable for plant growth. The use of some form of irrigation is

TABLE 1. Example of consumption of water by various crops, in inches*

Crop	April	May	June	July	Aug.	Sept.	Oct.	Seasonal total
Alfalfa	3.3	6.7	5.4	7.8	4.2	5.6	4.4	37.4
Beets		1.9	3.3	5.3	6.9	5.8	1.1	24.3
Cotton	1.1	2.0	4.1	5.8	8.6	6.7	2.7	31.0
Peaches	1.0	3.4	6.7	8.4	6.4	3.1	1.1	30.0
Potatoes			0.7	3.4	5.8	4.4		14.0

*1 in. = 25 mm.

well documented throughout the history of civilization.

Use of water by plants. Growing plants use water almost continuously. Growth of crops under irrigation is stimulated by optimum moisture, but retarded by excessive or deficient amounts. Factors influencing the rate of water use by plants include the type of plant and stage of growth, temperature, wind velocity, humidity, sunlight duration and intensity, and available water supply. Plants use the least amount of water upon emergence from the soil and near the end of the growing period. Irrigation and other management practices should be coordinated with the various stages of growth. A vast amount of research has been done on the use of water by plants, and results are available for crops under varying conditions. *See* PLANT GROWTH.

Consumptive use. In planning new or rehabilitating old irrigation projects, consumptive use is the most important factor in determining the amount of water required. It is also used to determine water rights.

Consumptive use, or evapotranspiration, is defined as water entering plant roots to build plant tissues, water retained by the plant, water transpired by leaves into the atmosphere, and water evaporated from plant leaves and from adjacent soil surfaces. Consumptive use of water by various crops under varying conditions has been determined by soil-moisture studies and computed by other well-established methods for many regions of the United States and other countries. Factors which have been shown to influence consumptive use are precipitation, air temperature, humidity, wind movement, the growing season, and latitude, which influences hours of daylight. **Table 1** shows how consumptive use varies at one location during the growing season. When consumptive-use data are computed for an extensive area, such as an irrigation project, the results will be given in acre-feet per acre for each month of the growing season and the entire irrigation period. Peak-use months determine system capacity needs. An acre-foot is the amount of water required to cover 1 acre 1 ft deep (approx. 44 ft^3 or 1214 m^3 of water).

Soil, plant, and water relationships. Soil of root-zone depth is the storage reservoir from which plants obtain moisture to sustain growth. Plants take from the soil not only water but dissolved minerals necessary to build plant cells. How often this reservoir must be filled by irrigation is determined by the storage capacity of the soil, depth of the root zone, water use by the crop, and the amount of depletion allowed before a reduction in yield or quality occurs. **Table 2** shows the approximate amounts of water held by soils of various textures.

Water enters coarse, sandy soils quite readily, but in heavy-textured soils the entry rate is slower. Compaction and surface conditions also affect the rate of entry.

Soil conditions, position of the water table, length of growing season, irrigation frequency, and other factors exert strong influence on root-zone depth. **Table 3** shows typical root-zone depths in well-drained, uniform soils under irrigation. The depth of rooting of annual crops increases during the entire growing period, given a favorable, unrestricted root zone. Plants in deep, uniform soils usually consume water more slowly from the lower root-zone area than from the upper. Thus, the upper portion is the first to be exhausted of moisture. For most crops, the entire root zone should be supplied with moisture when needed.

Maximum production can usually be obtained with most irrigated crops if not more than 50% of the available water in the root zone is exhausted during the critical stages of growth. Many factors influence this safe-removal percentage, including the type of

TABLE 2. Approximate amounts of water in soils available to plants

Soil texture	Water capacity in inches for each foot of depth (in centimeters for each meter of depth)
Coarse sandy soil	0.5–0.75 (4.15–6.25)
Sandy loam	1.25–1.75 (10.40–14.60)
Silt loam	1.75–2.50 (14.60–20.85)
Heavy clay	1.75–2.0 (14.60–16.65) or more

TABLE 3. Approximate effective root-zone depths for various crops

Crop	Root-zone depth, ft (m)
Alfalfa	6 (1.8)
Corn	3 (0.9)
Cotton	4 (1.2)
Potatoes	2 (0.6)
Grasses	2 (0.6)

crop grown and the rate at which water is being removed. Application of irrigation water should not be delayed until plants signal a need for moisture; wilting in the hot parts of the day may reduce crop yields considerably. Determination of the amount of water in the root zone can be done by laboratory methods, which are slow and costly. However, in modern irrigation practice, soil-moisture-sensing devices are used to make rapid determinations directly with enough accuracy for practical use. These devices, placed in selected field locations, permit an operator to schedule periods of water application for best results. Evaporation pans and weather records can be used to estimate plant-water use. Computerizing these data also helps farmers schedule their irrigations. The irrigation system should be designed to supply sufficient water to care for periods of most rapid evapotranspiration. The rate of evapotranspiration may vary from 0 to 0.4 in. per day (10 mm per day) or more. *See* PLANT-WATER RELATIONS.

Water quality. All natural irrigation waters contain salts, but only occasionally are waters too saline for crop production when used properly. When more salt is applied through water and fertilizer than is removed by leaching, a salt buildup can occur. If the salts are mainly calcium and magnesium, the soils become saline, but if salts predominantly are sodium, a sodic condition is possible. These soils are usually found in arid areas, especially in those areas where drainage is poor. Rainfall in humid areas usually carries salts downward to the ground water and eventually to the sea.

Saline soils may reduce yields and can be especially harmful during germination. Some salts are toxic to certain crops, especially when applied by sprinkling and allowed to accumulate on the plants. Salt levels in the soil can be controlled by drainage, by overirrigation, or by maintaining a high moisture level which keeps the salts diluted. *See* PLANTS OF SALINE ENVIRONMENTS.

Sodic soils make tillage and water penetration difficult. Drainage, addition of gypsum or sulfur, and overirrigation usually increase productivity.

Ponding or sprinkling can be used to leach salts. Intermittent application is usually better and, when careful soil-moisture management is practiced, only small amounts of excess irrigation are needed to maintain healthy salt levels.

Diagnoses of both water and soil are necessary for making management decisions. Commercial laboratories and many state universities test both water and soil, and make recommendations.

Methods of application. Water is applied to crops by surface, subsurface, sprinkler, and drip irrigation. Surface irrigation includes furrow and flood methods.

Furrow method. This method is used for row crops (**Fig. 1**). Corrugations or rills are small furrows used on close-growing crops. The flow, carried in furrows, percolates into the soil. Flow to the furrow is usually supplied by siphon tubes, spiles, gated pipe, or valves from buried pipe. Length of furrows and size of stream depend on slope, soil type, and crop; infiltration and erosion must be considered.

Fig. 1. Furrow method of irrigation. Water is supplied by pipes with individual outlets, or by ditches and siphon tubes.

Flood method. Controlled flooding is done with border strips, contour or bench borders, and basins. Border strip irrigation is accomplished by advancing a sheet of water down a long, narrow area between low ridges called borders. Moisture enters the soil as the sheet advances. Strips vary from about 20 to 100 ft (6 to 30 m) in width, depending mainly on slope (both down and across), and amount of water available. The border must be well leveled and the grade uniform; best results are obtained on slopes of 0.5% or less. The flood method is sometimes used on steeper slopes, but maldistribution and erosion make it less effective.

Fig. 2. A side-roll sprinkler system which uses the main supply line (often more than 1000 ft or 300 m long) to carry the sprinkler heads and as the axle for wheels.

Fig. 3. Center-pivot systems are very popular in new irrigation developments.

Bench-border irrigation is sometimes used on moderately gentle, uniform slopes. The border strips, instead of running down the slope, are constructed across it. Since each strip must be level in width, considerable earth moving may be necessary.

Basin irrigation is well adapted to flatlands. It is done by flooding a diked area to a predetermined depth and allowing the water to enter the soil throughout the root zone. Basin irrigation may be utilized for all types of crops, including orchards where soil and topographic conditions permit.

Subirrigation. This type of irrigation is accomplished by raising the water table to the root zone of the crop or by carrying moisture to the root zone by perforated underground pipe. Either method requires special soil conditions for successful operation.

Fig. 4. Center-pivot irrigation system irrigating corn. A single unit can water as much as 500 acres (200 hectares) or more. (*Eastern Iowa Light and Power Cooperative*)

Sprinkler systems. A sprinkler system consists of pipelines which carry water under pressure from a pump or elevated source to lateral lines along which sprinkler heads are spaced at appropriate intervals. Laterals are moved from one location to another by hand or tractor, or they are moved automatically. The side-roll wheel system, which utilizes the lateral as an axle (**Fig. 2**), is very popular as a labor-saving method. The center-pivot sprinkler system (**Fig. 3**) consists of a lateral carrying the sprinkler heads, and is moved by electrical or hydraulic power in a circular course irrigating an area containing up to 135–145 acres (546,200–586,700 m^2).

Extra equipment can be attached in order to irrigate the corners, or solid sets can be used. Solid-set systems are systems with sufficient laterals and sprinklers to irrigate the entire field without being moved. These systems are quite popular for irrigating vegetable crops or other crops requiring light, frequent irrigations and, in orchards, where it is difficult to move the laterals.

Sprinkler irrigation has the advantage of being adaptable to soils too porous for other systems. It can be used on land where soil or topographic conditions are unsuitable for surface methods. It can be used on steep slopes and operates efficiently with a small water supply.

Drip irrigation. This is a method of providing water to plants almost continuously through small-diameter tubes and emitters. It has the advantage of maintaining high moisture levels at relatively low capital costs. It can be used on very steep, sandy, and rocky areas and can utilize saline waters better than most other systems. Clean water, usually filtered, is necessary to prevent blockage of tubes and emitters. The system has been most popular in orchards and vineyards, but is also used for vegetables, ornamentals, and landscape plantings.

Automated systems. Automation is being used with solid-set and continuous-move types of systems, such as the center-pivot (**Fig. 4**) and lateral-move. Surface-irrigated systems are automated with check dams, operated by time clocks or volume meters, which open or close to divert water to other areas. Sprinkler systems, pumps, and check dams can all be activated by radio signals or low-voltage wired systems, which, in turn, can be triggered by soil-moisture-sensing devices or water levels in evaporation pans.

Automatically operated pumpback systems, consisting of a collecting pond and pump, are being used on surface-irrigated farms to better utilize water and prevent silt-laden waters from returning to natural streams.

Multiple uses. With well-designed and -managed irrigation systems, it is possible to apply chemicals and, for short periods of time, to moderate climate. Chemicals which are being used include fertilizers, herbicides, and some fungicides. Effectiveness depends on uniformity of mixing and distribution and on application at the proper times. Chemicals must be registered to be used in this manner.

Solid-set systems are frequently used to prevent frost damage to plants and trees, since, as water freezes, it releases some heat. A continuous supply of water is needed during the protecting period. However, large volumes of water are required, and ice loads may cause limb breakage. Sequencing of sprinklers for cooling ensures bloom delay in the spring and reduces heat damage in the summer.

Humid and arid regions. The percentage of increase in irrigated land is greater in humid areas than in arid and semiarid areas, although irrigation programs are often more satisfactory where the farmer does not depend on rainfall for crop growth. Good yields are obtained by well-timed irrigation, maintenance of high fertility, keeping the land well cultivated, and using superior crop varieties.

There is little difference in the principles of crop production under irrigation in humid and arid regions. The programming of water application is more difficult in humid areas because natural precipitation cannot be accurately predicted. Most humid areas utilize the sprinkler method.

To be successful, any irrigation system in any location must have careful planning with regard to soil conditions, topography, climate, cropping practices, water quality and supply, as well as engineering requirements. *See* AGRICULTURAL SOIL AND CROP PRACTICES; LAND DRAINAGE (AGRICULTURE); TERRACING (AGRICULTURE); WATER CONSERVATION.

Mel A. Hagood

Bibliography. R. Cuenca, *Irrigation System Design*, 1989; D. R. Hay (ed.), *Planning Now for Irrigation and Drainage in the 21st Century*, 1988; D. Hillel (ed.), *Advances in Irrigation Engineering*, vol. 4, 1987.

Isentropic flow

Compressible flow in which entropy remains constant along streamlines. Generally this implies that entropy is the same everywhere in the flow, in which case the flow is also referred to as homentropic flow. *See* COMPRESSIBLE FLOW; ENTROPY; ISENTROPIC PROCESS.

Because of the second law of thermodynamics, an isentropic flow does not strictly exist. From the definition of entropy, an isentropic flow is both adiabatic and reversible. However, all real flows experience to some extent the irreversible phenomena of friction, thermal conduction, and diffusion. Any nonequilibrium, chemically reacting flow is also irreversible. However, there are a large number of gas dynamic problems with entropy increase negligibly slight, which for the purpose of analysis are assumed to be isentropic. Examples are flow through subsonic and supersonic nozzles, as in wind tunnels and rocket engines; and shock-free flow over a wing, fuselage, or other aerodynamic shape. For these flows, except for the thin boundary-layer region adjacent to the surface where friction and thermal conduction effects can be strong, the outer inviscid flow can be considered isentropic. If shock waves exist in the flow, the entropy increase across these shocks destroys the assumption of isentropic flow, although the flow along streamlines between shocks may be isentropic. *See* ADIABATIC PROCESS; BOUNDARY-LAYER FLOW; NOZZLE; SHOCK WAVE; SUBSONIC FLIGHT; THERMODYNAMIC PRINCIPLES; THERMODYNAMIC PROCESSES.

The assumption of an isentropic flow greatly simplifies the analysis of a flowfield. In many cases, the simple statement of constant entropy replaces a complicated equation such as the momentum or energy equation. In particular, in an isentropic flow of a calorically perfect gas (constant specific heat), the pressure, density, and temperature of the flow are uniquely related to the local Mach number, as shown in Eqs. (1)–(3), where γ is the ratio of specific heat at constant pressure to specific heat at constant volume; p, ρ, and T are the local static pressure, density, and temperature, respectively; M is the local Mach number; and p_0, ρ_0, and T_0 are the total pressure, density, and temperature, respectively. (Total conditions are defined as the conditions that would exist at a point in the flow if the velocity at that point were isentropically reduced to zero.)

$$\frac{p_0}{p} = \left(1 + \frac{\gamma - 1}{2}M^2\right)^{\gamma/(\gamma-1)} \qquad (1)$$

$$\frac{\rho_0}{\rho} = \left(1 + \frac{\gamma - 1}{2}M^2\right)^{\gamma/(\gamma-1)} \qquad (2)$$

$$\frac{T_0}{T} = 1 + \frac{\gamma - 1}{2}M^2 \qquad (3)$$

Equation (3), involving T, holds for any adiabatic flow of a calorically perfect gas; such a flow does not have to be isentropic. Equations (1)–(3) are applicable to isentropic flows of any geometric complexity, including three-dimensional flows. In this manner, isentropic flows are much easier to compute than nonisentropic. *See* FLUID-FLOW PRINCIPLES; GAS DYNAMICS.

John D. Anderson, Jr.

Bibliography. J. D. Anderson, Jr., *Fundamentals of Aerodynamics*, 3d ed., 2001; J. D. Anderson, Jr., *Introduction to Flight*, 5th ed., 2005; J. D. Anderson, Jr., *Modern Compressible Flow*: *With Historical Perspective*, 3d ed., 2003.

Isentropic process

In thermodynamics a change that is accomplished without any increase or decrease of entropy is referred to as isentropic. Since the entropy always increases in a spontaneous process, one must consider reversible or quasistatic processes. During a reversible process the quantity of heat transferred, dQ, is directly proportional to the system's entropy change, dS, as in Eq. (1), where T is the absolute temperature of the system. Systems which are thermally

$$dQ = T\,dS \qquad (1)$$

insulated from their surroundings undergo processes without any heat transfer; such processes are called adiabatic. Thus during an isentropic process there are no dissipative effects and, from Eq. (1), the system neither absorbs nor gives off heat. For this reason the isentropic process is sometimes called the reversible adiabatic process. *See* ADIABATIC PROCESS; ENTROPY; THERMODYNAMIC PROCESSES.

Work done during an isentropic process is produced at the expense of the amount of internal energy stored in the nonflow or closed system. Thus, the useful expansion of a gas is accompanied by a marked decrease in temperature, tangibly demonstrating the decrease of internal energy stored in the system. For ideal gases the isentropic process can be expressed by Eq. (2), where P is the pressure, V is

$$P_1 V_1^\gamma = P_2 V_2^\gamma = \text{constant} \tag{2}$$

the volume, and γ is the ratio between the specific heat at constant pressure and the specific heat at constant volume for the given gas. It can be closely approximated by the values of 1.67 and 1.40 for dilute monatomic and diatomic gases, respectively. For a comparison of various processes involving a gas *see* POLYTROPIC PROCESS. Philip E. Bloomfield

Isentropic surfaces

Surfaces along which the entropy and potential temperature of air are constant. Potential temperature, in meteorological usage, is defined by the relationship given below,

$$\Theta = T\left(\frac{1000}{P}\right)^{(c_p - c_v)/c_p}$$

in which T is the air temperature, P is atmospheric pressure expressed in millibars, C_p is the heat capacity of air at constant pressure and C_v is the heat capacity at constant volume. Since the potential temperature of an air parcel does not change if the processes acting on it are adiabatic (no exchange of heat between the parcel and its environment), a surface of constant potential temperature is also a surface of constant entropy. The slope of isentropic surfaces in the atmosphere is of the order of $^1/_{100}$ to $^1/_{1000}$. An advantage of representing meteorological conditions on isentropic surfaces is that there is usually little air motion through such surfaces, since thermodynamic processes in the atmosphere are approximately adiabatic. *See* ADIABATIC PROCESS; ATMOSPHERIC GENERAL CIRCULATION. Frederick Sanders

Bibliography. J. R. Holton, *An Introduction to Dynamic Meteorology*, 4th ed., 2004; J. C. McWilliams, *Fundamentals of Geophysical Fluid Dynamics*, 2006; T. D. Potter and B. R. Colman (eds.), *Handbook of Weather, Climate and Water: Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements*, 2003.

Ising model

A model which consists of a lattice of "spin" variables with two characteristic properties: (1) each of the spin variables independently takes on either the value +1 or the value −1; and (2) only pairs of nearest neighboring spins can interact. The study of this model (introduced by Ernst Ising in 1925) in two dimensions, where many exact calculations have been carried out explicitly, forms the basis of the modern theory of phase transitions and, more generally, of cooperative phenomena.

The two-dimensional Ising model was shown to have a phase transition by R. E. Peierls in 1936, and the critical temperature or Curie temperature, that is, the temperature at which this phase transition takes place, was calculated by H. A. Kramers and G. H. Wannier and by E. W. Montroll in 1941. Major breakthroughs were accomplished by Lars Onsager in 1944, by Bruria Kaufman and Onsager in 1949, and by Chen Ning Yang in 1952. Onsager first obtained the free energy and showed that the specific heat diverges as $-\ln|1 - T/T_c|$ when the temperature T is near the critical temperature T_c; Kaufman and Onsager computed the short-range order; and Yang calculated the spontaneous magnetization. Since then several other properties have been obtained, and since 1974 connections with relativistic quantum field theory have been made. *See* QUANTUM FIELD THEORY.

Cooperative phenomena. A macroscopic piece of material consists of a large number of atoms, the number being of the order of Avogadro's number (approximately 6×10^{23}). Thermodynamic phenomena all depend on the participation of such a large number of atoms. Even though the fundamental interaction between atoms is short-ranged, the presence of this large number of atoms can, under suitable conditions, lead to an effective interaction between widely separated atoms. Phenomena due to such effective long-range interactions are referred to as cooperative phenomena. The simplest examples of cooperative phenomena are phase transitions. The most familiar phase transition is either the condensation of steam into water or the freezing of water into ice. Only slightly less familiar is the ferromagnetic phase transition that takes place at the Curie temperature, which, for example, is roughly 1043 K for iron.

Of the several models which exhibit a phase transition, the Ising model is the best known. In three dimensions the model is so complicated that no exact computation has ever been made, while in one dimension the Ising model does not undergo a phase transition. However, in two dimensions the Ising model not only has a ferromagnetic phase transition but also has very many physical properties which may be exactly computed. Indeed, despite the restriction on dimensionality, the two-dimensional Ising model exhibits all of the phenomena peculiar to magnetic systems near the Curie temperature. *See* CURIE TEMPERATURE; FERROMAGNETISM; MAGNETISM.

Definition of model. The mutual interaction energy of the pair of spins σ_α and $\sigma_{\alpha'}$ when α and α' are nearest neighbors may be written as $-E(\alpha,\alpha') \cdot \sigma_{\alpha'}$. The meaning of this is that the interaction energy is $-E(\alpha,\alpha')$ when σ_α and $\sigma\alpha'$ are both $+1$ or -1, and is $+E(\alpha,\alpha')$ when $\sigma_\alpha = 1$, $\sigma\alpha' = -1$, or $\sigma_\alpha = -1$, $\sigma\alpha' = 1$. In addition, a spin may interact with an external magnetic field H with energy $-H\sigma\alpha$. From these two basic interactions the total interaction energy for the square lattice may be written as Eq. (1), where j

$$E = -\sum_j \sum_k [E_1(j,k)\sigma_{j,k}\sigma_{j,k+1} + E_2(j,k)\sigma_{j,k}\sigma_{j+1,k} + H\sigma_{j,k}] \quad (1)$$

specifies the row and k specifies the column of the lattice. In this form the interaction energies $E_1(j,k)$ and $E_2(j,k)$ are allowed to vary arbitrarily throughout the lattice. A special case of great importance is the translationality invariant case (E_1 and E_2 independent of j and k) which was studied by Onsager in 1944. This is the model needed to study a pure ferromagnet without impurities.

Several generalizations of Ising's original model have been considered. For example, σ can be allowed to take on more values than just ±, and interactions other than nearest neighbor can be used. For these generalizations no exact calculations have been performed in two or three dimensions. However, various approximate calculations indicate that the phase transition properties of these models are the same as those of the Onsager lattice.

The extension to the nontranslationally invariant case where $E_1(j,k)$ and $E_2(j,k)$ are treated as independent random variables is important for studying the effects of impurities in ferromagnetics.

Thermodynamic properties. The basic simplification in framing the definition of the Ising model of the preceding section is the choosing of the fundamental variables to be the numbers $\sigma_{j,k}$ which can be only $+1$ or -1. Because of this choice there can be no terms in the interaction energy which refer to kinetic energy or to angular momentum. Consequently, the $\sigma_{j,k}$ do not change with time, and study of the system is, by necessity, confined to those physical properties which depend only on the distribution of energy levels of the system. When the number of energy levels is large, this study requires the use of statistical mechanics.

Statistical mechanics allows the calculation of average macroscopic properties from the microscopic interaction E. If A is some property of the spins σ of the system, then the thermodynamic average of A is given by Eq. (2), where T is the temperature, k is Boltzmann's constant, Z is given by Eq. (3),

$$\langle A \rangle \lim_{N\to\infty} \frac{1}{Z} \sum_{\{\sigma\}} A e^{-E(\sigma)/kT} \quad (2)$$

$$Z = \sum e^{-E(\sigma)/kT} \quad (3)$$

the sums are over all values of $\sigma_{j,k} = \pm 1$, and N is the number of rows and the number of columns. It is mandatory that the thermodynamic limit $N \to \infty$ be taken for these thermodynamic averages to have a precise meaning. *See* BOLTZMANN CONSTANT.

The most important thermodynamic properties of a ferromagnet are the internal energy per site $u = \langle E/N^2 \rangle$ the specific heat $c = \partial u / \partial T$, the magnetization per site $M = \langle \sigma \rangle$ and the magnetic susceptibility $\chi = \partial M \partial H$. These quantities have been computed for the two-dimensional Ising model $E_1 \neq E_2$, but for convenience this discussion is restricted to $E_1 = E_2 = E$.

Onsager studied the two-dimensional square lattice at $H = O$ and computed the specific heat exactly. From that calculation he found that the specific heat was infinite at the critical temperature of Kramers and Wannier given as the solution of Eq. (4). When

$$\sinh \frac{2E}{kT} = 1 \quad (4)$$

T is close to the critical temperature T_c, the specific heat is approximated by Eq. (5). The behavior of

$$c \sim -\frac{8E^2}{k{T_c}^2\pi} \ln \left| 1 - \frac{T}{T_c} \right| \quad (5)$$

the specific heat for any temperature is plotted in **Fig. 1**.

The spontaneous magnetization isdefined as

$$M(0) = \lim_{H\to 0^+} M(H)$$

For $T > T_c$, $M(0) = 0$. For $T < T_c$, Yang found that $M(0)$ is given by Eq. (6). When T is near T_c, $M(0)$ is approximated by Eq. (7). The behavior $M(0)$ as a function of T is plotted in **Fig. 2**.

$$M(0) = \left[1 - \left(\sinh \frac{2E}{kT} \right)^{-4} \right]^{1/8} \quad (6)$$

$$M(0) \sim \left[\frac{8\sqrt{2}}{kT_c} \left(1 - \frac{T}{T_c} \right) \right]^{1/8} \quad (7)$$

The magnetic susceptibility χ at $H = 0$ is much

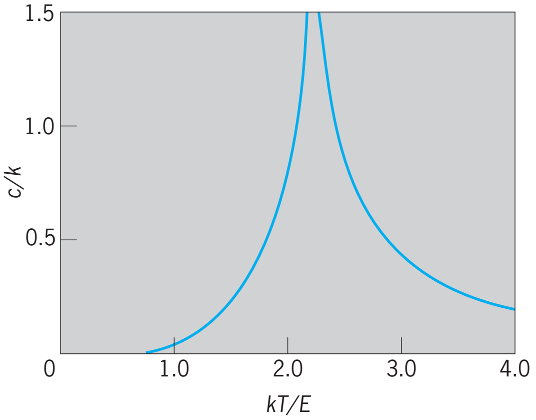


Fig. 1. Specific heat of Onsager's lattice for $E_2/E_1 = 1$.

more difficult to compute than either the specific heat or the spontaneous magnetization. Indeed, no closed form expression is known for χ over the entire range of temperature. However, near T_c it is known that as $T \to T_c^+$, χ is approximated by Eq. (8), and, as $T \to T_c^-$, is approximated by Eq. (9).

$$\chi(T) \sim C_0^+ \left|1 - \frac{T_c}{T}\right|^{-7/4} + C_1^+ \left|1 - \frac{T_c}{T}\right|^{-3/4} + C_2 \qquad (8)$$

$$\chi(T) \sim C_0^- \left|1 - \frac{T_c}{T}\right|^{-7/4} + C_1^- \left|1 - \frac{T_c}{T}\right|^{-3/4} + C_2 \qquad (9)$$

where

$$\begin{aligned} C_0^+ &= 0.96258\,17322\cdots \\ C_0^- &= 0.02553\,69719\cdots \\ C_1^+ &= 0.07498\,81538\cdots \\ C_1^- &= -0.0198\,94107\cdots \end{aligned}$$

See INTERNAL ENERGY; MAGNETIC SUSCEPTIBILITY; MAGNETIZATION; SPECIFIC HEAT OF SOLIDS; STATISTICAL MECHANICS; THERMODYNAMIC PRINCIPLES.

Random impurities. A question which can be very usefully studied in the Ising model is the generalization of statistical mechanics to deal with the experimental situation in which the interaction energy of the system is not completely known because of the presence of impurities. The term "impurity" refers not only to the presence of foreign material in a sample but to any physical property such as defects or isotopic composition which makes lattice sites different from one another. The distribution of these impurities is governed by spin-independent forces. At least two different situations can be distinguished. (1) As the temperature changes, the distribution of impurities may change; such a situation will occur, for example, near the melting point of a lattice. (2) The distribution of impurities may be independent of temperature, at least on the time scale of laboratory measurements; such a distribution will obtain when the temperature of a lattice is well below the melting temperature. Impurities of this sort are said to be frozen in. *See* CRYSTAL DEFECTS.

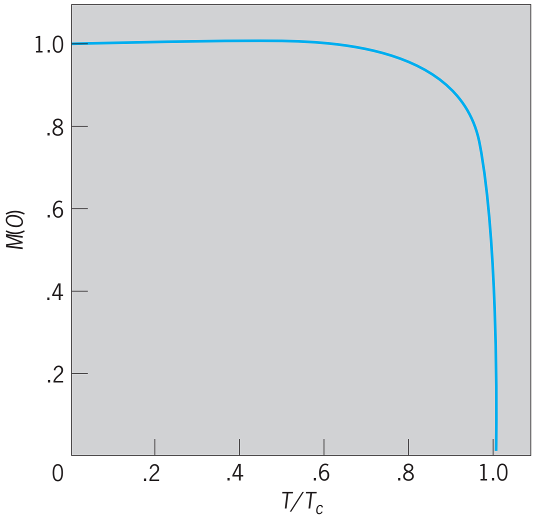


Fig. 2. Spontaneous magnetization *M*(0) of Onsager's lattice for $E_2/E_1 = 1$.

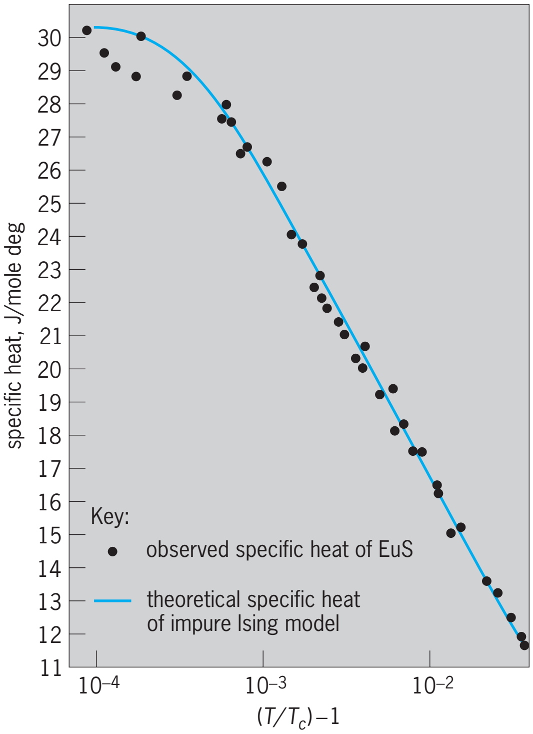


Fig. 3. Comparison of the impure Ising model specific heat with the observed specific heat of EuS for $T > T_c$.

For a study of frozen-in impurities to be realistic, the impurities must be distributed at random throughout the lattice. The translational invariance of the system is now totally destroyed, and it is not at all clear that the phase transition behavior of the pure and random system should be related to each other at all.

These problems were studied for a special case of the Ising model by McCoy and Wu in 1968. They let $E_2(j, k)$ depend on j but not k and kept $E_1(j, k)$ independent of both j and k. Then the variables $E_2(j)$ were chosen with a probability distribution $P(E_2)$. When $P(E_2)$ was of narrow width, they showed that logarithmic divergence of Onsager's specific heat is smoothed out into an infinitely differentiable essential singularity. Such a smoothing out of sharp phase transition behavior may in fact have been observed. The results of one such experiment, carried out by B. J. C. van der Hoeven and colleagues, are compared with the results of the Ising model random specific heat calculation in **Fig. 3**.

Barry M. McCoy; Tai Tsun Wu

Bibliography. C. Domb and M. Green, *Phase Transitions and Critical Phenomena*, 1972; R. Liebmann, *Statistical Mechanics of Periodic Frustrated Ising Systems*, 1986; B. McCoy and T. T. Wu, *The Two Dimensional Ising Model*, 1973.

Island biogeography

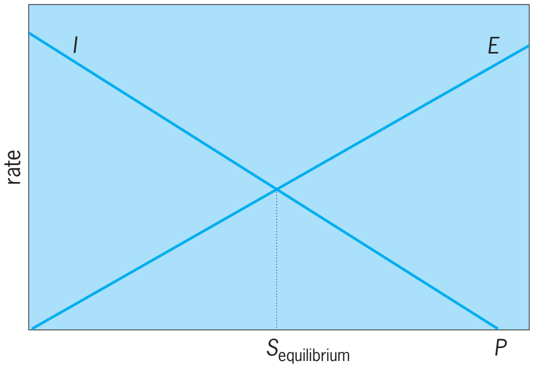
The distribution of plants and animals on islands. Islands harbor the greatest number of endemic species. The relative isolation of many islands has allowed populations to evolve in the absence of competitors and predators, leading to the evolution of unique species that can differ dramatically from their mainland ancestors.

Dispersal of species. Plant species produce seeds, spores, and fruits that are carried by wind or water currents, or by the feet, feathers, and digestive tracts of birds and other animals. The dispersal of animal species is more improbable, but animals can also be carried long distances by wind and water currents, or rafted on vegetation and oceanic debris. Long-distance dispersal acts as a selective filter that determines the initial composition of an island community. For example, land snails of many tropical Pacific islands are dominated by small-bodied species that have probably been carried by wind dispersal. Many species of continental origin may never reach islands unless humans accidentally or deliberately introduce them. Consequently, although islands harbor the greatest number of unique species, the density of species on islands (number of species per area) is typically lower than the density of species in mainland areas of comparable habitat. *See* POPULATION DISPERSAL.

Unique morphological traits. Once a species reaches an island and establishes a viable population, it may undergo evolutionary change because of genetic drift, climatic differences between the mainland and the island, or the absence of predators and competitors from the mainland. Consequently, body size, coloration, and morphology of island species often evolve rapidly, producing forms unlike any related species elsewhere. Examples include the giant land tortoises of the Galápagos, and the Komodo dragon, a species of monitor lizard from Indonesia. *See* POLYMORPHISM (GENETICS); POPULATION GENETICS; SQUAMATA.

If enough morphological change occurs, the island population becomes reproductively isolated from its mainland ancestor, and it is recognized as a unique species. Because long-distance dispersal is relatively infrequent, repeated speciation may occur as populations of the same species successively colonize an island and differentiate. The most celebrated example is Darwin's finches, a group of related species that inhabit the Galápagos Islands and were derived from South American ancestors. The island species have evolved different body and bill sizes, and in some cases occupy unique ecological niches that are normally filled by mainland bird species such as warblers and woodpeckers. The morphology of these finches was first studied by Charles Darwin and constituted important evidence for his theory of natural selection. *See* ANIMAL EVOLUTION; SPECIATION.

Theory. For most groups of organisms, the number of species increases in a nonlinear fashion with island area. Mathematically, the relationship can be described by a power function, $S = kA^z$, where S is the number of species on the island, A is the island area, and k and z are fitted constants. On a double logarithmic plot of species versus area, the relationship is linear, with the constant z describing the slope of the line. Although the slope of the species-area curve is steeper for isolated archipelagoes, the number of species per unit area tends to decrease with increasing isolation.



Equilibrium number of species ($S_{\text{equilibrium}}$) is determined by the intersection of the immigration (*I*) and extinction (*E*) rate curves. *P* is the maximum number of species in the source pool.

Although descriptions of the effects of area and isolation on species richness began about a century ago, only since the late 1960s have these observations been incorporated into a quantitative model. The MacArthur-Wilson equilibrium theory of island biogeography proposes that the number of species on an island represents an ecological balance between ongoing colonization and extinction of populations (see **illus.**). The model assumes that an island receives its colonists from a permanent source pool (P) of mainland species. The colonization rate (number of new species per unit time) is at a maximum value (I) when the island is empty, and the extinction rate (number of established island populations going extinct per unit time) is at a maximum value (E) when all P species are present on the island. The equilibrium number of species is determined by the intersection of the colonization and extinction curves, which occurs at the point $S_{\text{equilibrium}} = PI/(I + E)$. The model assumes that population size is proportional to island area, and that the risk of population extinction increases at small population size. Consequently, the extinction curve for small islands is steeper than for large islands, producing a trend toward more species at equilibrium as island size increases. The model also assumes that the immigration curve is shallower for more isolated islands, reducing the equilibrium number of species.

The key prediction of the model is that island populations are transient and that colonizations and extinctions are frequent. The most successful test of the equilibrium theory was carried out on tiny mangrove islands in the Florida Keys that were experimentally defaunated and then recolonized by insects over a one-year period. The experiment confirmed many predictions of the equilibrium model, although much of the turnover may have been from transient species that had not established viable populations.

Nature reserves. The equilibrium theory of island biogeography has been used as a tool for guiding conservation decisions. Its principal prescription is that large, contiguous areas will conserve the maximum number of species. However, specific application of the model has been problematic because many plant and animal communities do not exhibit the turnover predicted by the model. Moreover, the model predicts only the number of species, not their identities. Finally, the model assumes an intact mainland fauna, which may no longer exist in fragmented landscapes. *See* LANDSCAPE ECOLOGY; ZOOGEOGRAPHY.

Advances. Island biogeography theory has recently been extended to describe the persistence of single-species metapopulations. A metapopulation is a set of connected local populations in a fragmented landscape that does not include a persistent source pool region. Instead, the fragments themselves serve as stepping stones for local colonization and extinction. The most successful application of the metapopulation model has been to spotted owl populations of old-growth forest fragments in the northwestern United States. The model prescribes the minimum number, area, and spatial arrangement of forest fragments that are necessary to ensure the long-term persistence of the spotted owl metapopulation. *See* BIOGEOGRAPHY; ECOLOGICAL COMMUNITIES; ECOSYSTEM.

Nicholas J. Gottelli

Bibliography. T. J. Case and M. L. Cody, Testing theories of island biogeography, *Amer. Sci.*, 75:402–411, 1987; R. H. MacArthur and E. O. Wilson, *The Theory of Island Biogeography*, Princeton University Press, Princeton, NJ, 1969; K. A. McGuinness, Equations and explanations in the study of species-area curves, *Biol. Rev.*, 59:423–440, 1984; D. Quammen, *The Song of the Dodo: Island Biogeography in an Age of Extinctions*, Scribner, New York, 1996; M. Williamson, *Island populations*, Oxford University Press, Oxford, 1981.

Isoantigen

An immunologically active protein or polysaccharide present in some but not all individuals in a particular species. These substances initiate the formation of antibodies when introduced into other individuals of the species that genetically lack the isoantigen. Like all antigens, they are also active in stimulating antibody production in heterologous species. The ABO, MN, and Rh blood factors in humans constitute important examples; thus, elaborate precautions for typing are required in blood transfusions. *See* BLOOD GROUPS.

Analogous situations exist for the bloods of most other animal species. Isoantigens are also believed responsible for the ultimate failure of tissue grafts between individuals of the same species, except that of the same genetic constitution or those that have been rendered tolerant. *See* ACQUIRED IMMUNOLOGICAL TOLERANCE; TRANSPLANTATION BIOLOGY.

Isoantigens are to be distinguished from autoantigens, which are antigens active even in the species from which they are derived and in individuals who already possess the antigen. Brain and lens tissue, as well as sperm, constitute examples. These exceptions to the usual rule of nonantigenicity for self-constituents may be more apparent than real, however, since the substances cited are all protected to some degree from contact with the blood, and thus normally do not reach the sites of antibody formation except after experimental manipulation.

Autoantibodies may also be produced in various disease states, perhaps as a result of modification of normal host tissue by the infecting microorganism or by altered host metabolism. Examples are the paroxysmal hemoglobinuria observed in syphilis, acquired hemolytic anemia, or some of the manifestations in rheumatic fever. *See* ANTIBODY; ANTIGEN; AUTOIMMUNITY; POLYSACCHARIDE; PROTEIN.

Margaret J. Polley

Bibliography. E. L. Cooper et al. (eds.), *Developmental and Comparative Immunology*, 1987; D. Male et al., *Advanced Immunology*, 1987; S. B. Mizel and P. Jaret, *The Human Immune System: The New Frontier in Medicine*, 1986; P. N. Plowman, *Hematology and Immunology*, 1987.

Isobar (meteorology)

A curve along which pressure is constant. Leading examples of its uses are in weather forecasting and meteorology. The most common weather maps are charts of weather conditions at the Earth's surface and mean sea level, and they contain isobars as principal information. Areas of bad or unsettled weather are readily defined by roughly circular isobars around low-pressure centers at mean sea level. Likewise, closed isobars around high-pressure centers define areas of generally fair weather **(Fig. 1)**. *See* AIR PRESSURE.

Geostrophic wind. A principal use of isobars stems from the so-called geostrophic wind, which approximates the actual wind on a large scale. The direction of the geostrophic wind is parallel to the isobars, in the sense that if an observer stands facing away from the wind, higher pressures are to the person's right if in the Northern Hemisphere and to the left if in the Southern. Thus, in the Northern Hemisphere, flow is counterclockwise about low-pressure centers and clockwise about high-pressure centers, with the direction of the flow reversed in the Southern Hemisphere.

The speed of the geostrophic wind is inversely proportional to distance between isobars drawn at regular close intervals. More precisely, the geostrophic wind is as in Eq. (1).

$$V = \frac{1}{f\rho}\frac{\partial p}{\partial n} \qquad (1)$$

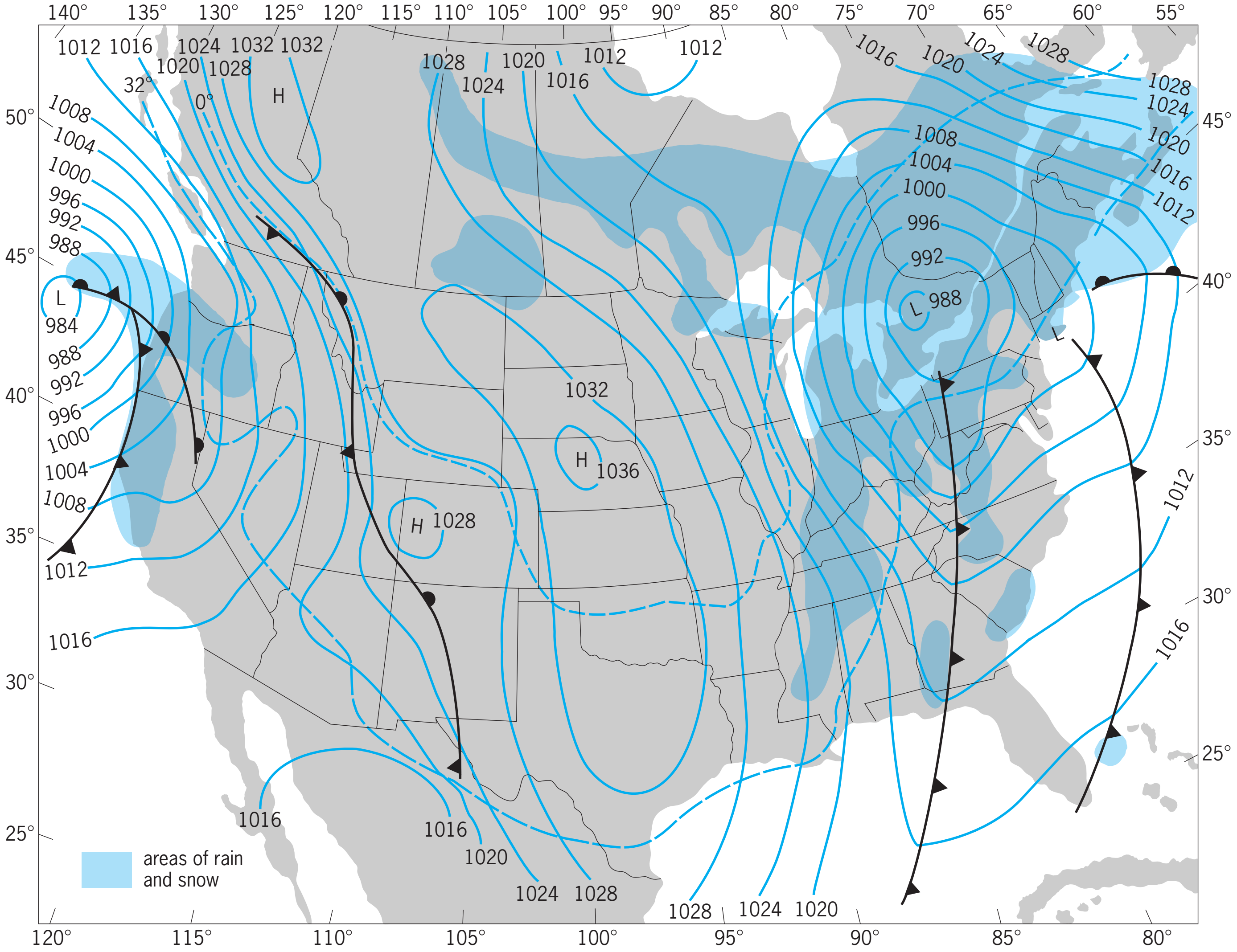


Fig. 1. Weather map at mean sea level for 7:00 A.M., Eastern Standard Time, January 14, 1979. Isobars are labeled in millibars. 1 mb = 10^2 Pa. (*After U.S. Department of Commerce, National Oceanic and Atmospheric Administration*, *Environmental Data and Information Service*, *Daily Weather Maps*, *Weekly Series*, *January 8–14, 1979*)

V = geostrophic wind speed
$f = 2\omega \sin \phi$, and is called the Coriolis parameter
$\omega = 7.9292 \times 10^{-5}$ radian s^{-1}, the angular velocity of the Earth
ϕ = latitude, positive in the Northern Hemisphere, negative in the Southern
ρ = density of the air, varying about a value of of 1.2×10^{-3} g cm^{-3} at mean sea level
p = pressure, varying about a value of 1013 millibars (101.3 kPa) at mean sea level
n = distance normal to isobar and toward higher pressure

Note that f is positive in the Northern Hemisphere and negative in the Southern. The geostrophic wind law written this way, then, provides a single rule for both Northern and Southern hemispheres, if it is understood that a change in sign of V implies reversal in direction.

When the geostrophic wind is calculated from isobars on a weather map, Eq. (1) is approximated by Eq. (2) [**Fig. 2**], where Δp is the interval of pressure

$$V = \frac{1}{f\rho}\,\frac{\Delta p}{d} \tag{2}$$

at which the isobars are drawn, and d is the distance between two adjacent isobars. Thus, if isobars are drawn at 4-millibar intervals, then $\Delta p = 4000$ g cm^{-1} s^{-2}; and if $d = 250$ km $= 2.5 \times 10^7$ cm, $f = 10^{-4}$ radian s^{-1} ($\phi = 43.3$°N), and $\rho = 1.2 \times 10^{-3}$ g cm^{-3}, then

$$\begin{aligned} V &= \frac{4000}{10^{-4} \times 1.2 \times 10^{-3} \times 2.5 \times 10^{7}}\ \text{cm s}^{-1} \\ &= 1333\ \text{cm s}^{-1} = 13\ \text{m s}^{-1} = 25\ \text{knots} \end{aligned}$$

Although the geostrophic wind is useful everywhere as an approximation to the wind (except in low latitudes where f approaches zero), at mean sea level it is only a very rough approximation. The departure of the actual wind at mean sea level from

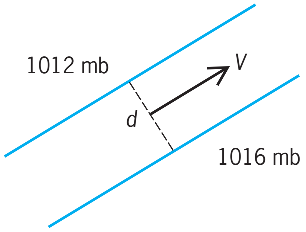


Fig. 2. Schematic of two isobars illustrating the elements in Eq. (2).

the geostrophic is due principally to friction and turbulence, which are highly variable; and also to the impossibility of precisely reducing pressure to mean sea level from plateaus such as the Rocky Mountains and Himalayas. Thus, although the speed of the geostrophic wind varies with air density, at mean sea level it is hardly worthwhile to take into account the variations of air density. A constant, such as $\rho = 1.2 \times 10^{-3}$ g cm^{-3}, is often used instead.

In the free atmosphere, however, where modern aircraft fly, the geostrophic wind is quite accurate as an approximation to the actual wind. Above 1 km (0.6 mi) or so, turbulence and frictional forces are very small, and at even greater heights, above Earth's plateaus, the pressure reduction problem vanishes. In the upper atmosphere, therefore, it is worthwhile to take into account the 15% or so variation of density in the horizontal. *See* GEOSTROPHIC WIND.

Isobaric surfaces. In practice, density variations in the free atmosphere are rather neatly taken into account by constructing weather charts on isobaric surfaces (surfaces of constant pressure), rather than on level surfaces (surfaces of constant height).

Isobars are meaningless on isobaric surfaces, since pressure is constant along any arbitrary curve on such a surface. Instead, curves of constant height are used. These are called isohyets or, more commonly, height contours. The geostrophic wind on isobaric surfaces is as in Eq. (3), where z is height, g is force

$$V = \frac{g}{f}\frac{\partial z}{\partial n} \tag{3}$$

of gravity, and n is distance normal to the contour in the direction of increasing height. *See* WEATHER MAP.

Frederick G. Shuman

Isobar (nuclear physics)

One of two or more atoms which have a common mass number A but which differ in atomic number Z. Thus, although isobars possess approximately equal masses, they differ in chemical properties; they are atoms of different elements. Isobars whose atomic numbers differ by unity cannot both be stable; one will inevitably decay into the other by β^- emission ($Z \rightarrow Z + 1$), β^+ emission ($Z \rightarrow Z - 1$), or electron capture ($Z \rightarrow Z - 1$). There are many examples of stable isobaric pairs, such as ^{50}Ti ($Z = 24$) and ^{50}Cr ($Z = 26$), and four examples of stable isobaric triplets. At most values of A the number of known radioactive isobars exceeds the number of stable ones. *See* ELECTRON CAPTURE; RADIOACTIVITY.

Henry E. Duckworth

Isobaric process

A thermodynamic process during which the pressure remains constant. When heat is transferred to or from a gaseous system, a volume change occurs at constant pressure. This thermodynamic process can be illustrated by the expansion of a substance when it is heated. The system is then capable of doing an amount of work on its surroundings. The maximum work is done when the external pressure P_{ext} of the surroundings on the system is equal to P, the pressure of the system. If V is the volume of the system, the work performed as the system moves from state 1 to 2 during an isobaric thermodynamic process, W_{12}, is the maximum work as given by Eq. (1). For

$$W_{12} = \int_1^2 P\,dV = P\int_1^2 dV = P(V_2 - V_1) \tag{1}$$

an ideal gas the volume increase is a result of a corresponding temperature increase, so that Eq. (1) yields Eq. (2), where n is the number of moles of gas and

$$W_{12} = nR(T_2 - T_1) \tag{2}$$

R is the gas constant. Then W_{12} represents the work done by the system and is positive if $V_2 > V_1$ (that is, $T_2 > T_1$ for an ideal gas). *See* HEAT TRANSFER.

By the first law of thermodynamics, the change of the internal energy, U, in any process is equal to the difference between the heat gained, Q_p, and the work done by the system, W_{12}; thus, Eq. (3) holds. From Eqs. (1) and (3), it follows that Eq. (4) holds for isobaric processes.

$$U_2 - U_1 = U_{12} = Q_p - W_{12} \tag{3}$$

$$Q_p = (U_2 + PV_2) - (U_1 + PV_1) \tag{4}$$

For isobaric processes, it is useful to introduce the enthalpy, H, which is given by the sum of the internal energy U and PV. Then Eq. (5) can be formulated to

$$Q_p = H_2 - H_1 = \int_1^2 C_p\,dT \tag{5}$$

represent Q_p, the transferred heat at constant pressure, where $C_p = (dQ/dT)|_p$ is the heat capacity at constant pressure. For an ideal gas, Eq. (6) holds.

$$Q_p = C_p(T_2 - T_1) \tag{6}$$

See ENTHALPY.

If the isobaric process is also a reversible process, Eq. (7) is obtained, where S is the entropy.

$$Q_p = \int_1^2 T\,dS \tag{7}$$

See ENTROPY.

For an ideal gas the internal energy change can be expressed in terms of the heat capacity at constant volume, as in Eq. (8). The combination of Eqs. (2), (3), (6), and (8) yields Eq. (9).

$$U_{12} = C_V(T_2 - T_1) \quad (8)$$

$$C_p = C_V + nR \quad (9)$$

From Eq. (8) it follows that, for an isometric process (fixed volume), heat goes into internal energy only. However, from Eqs. (9) and (3) it follows that, for an isobaric process, $C_p > C_V$, and the heat input is converted into internal energy increase as well as work output by the system. *See* ADIABATIC PROCESS; ISOMETRIC PROCESS; THERMODYNAMIC PROCESSES.

For a comparison of the isobaric process with other processes involving a gas *See* POLYTROPIC PROCESS.

Philip E. Bloomfield

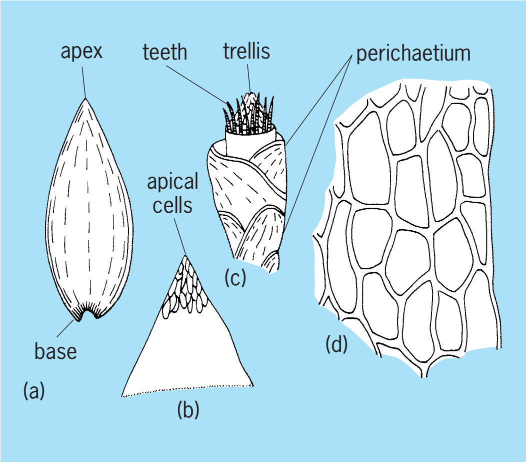
Bibliography. D. Halliday, R. Resnick, and J. Walker, *Fundamentals of Physics*, 6th ed., 2000; J. R. Howell and R. O. Buckius, *Fundamentals of Engineering Thermodynamics*, 2d ed., 1992; G. J. Van Wylen, R. E. Sonntag, and C. Borgnakke, *Fundamentals of Classical Thermodynamics*, 4th ed., 1994; M. W. Zemansky, M. M. Abbott, and H. C. Van Ness, *Basic Engineering Thermodynamics*, 2d ed., 1975.

Isobryales

An order of the true mosses (subclass Bryidae). This order is difficult to define precisely, but includes plants that generally grow from a creeping primary stem with leaves reduced or essentially lacking and plants that have spreading to ascending secondary stems which may be pinnately branched (see **illus.**). Paraphyllia and pseudoparaphyllia are sometimes present on the stems. The leaves may have single or double and sometimes short costae. The cells may be short or elongate and smooth or papillose, with those at basal angles sometimes differentiated. The sporophytes are lateral, usually with elongate setae and capsules. The double peristome, sometimes reduced, consists of 16 teeth which are papillose on the outer surface, or less often cross-striate at the base, and an endostome with narrow segments and a low basal membrane or none at all. The calyptrae are cucullate and naked, or mitrate and hairy

The order is surely heterogeneous in composition. It consists of about 19 families and 124 genera, some of which may be better placed in the Hypnales. *See* BRYIDAE; BRYOPHYTA; BRYOPSIDA; HYPNALES.

Howard Crum



***Fontinalis duriaei*, an example of the order Isobryales. (*a*) Leaf. (*b*) Apex of leaf. (*c*) Trellis and teeth. (*d*) Alar cells. (*After W. H. Welch, Mosses of Indiana, Indiana Department of Conservation, 1957*)**

Isocyanate

A derivative of isocyanic acid. Isocyanates are represented by the general formula —N=C=O, where R is predominantly alkyl or aryl; however, stable isocyanates in which the N=C=O group is linked to elements such as sulfur, silicon, phosphorus, nitrogen, or the halogens have also been prepared. Most members of this class of compounds are liquids that are sensitive to hydrolysis and are strong lacrimators. Isocyanates are extremely reactive, especially toward substrates containing active hydrogen. They have found wide use in the commercial manufacture of polyurethanes, which are used in making rigid and flexible foam, elastomers, coatings, and adhesives.

Preparation. Most isocyanates are prepared by reacting excess phosgene ($COCl_2$) with primary amines in high-boiling-point solvents, as in reaction (1). The reaction involves several steps: the

$$\mathrm{R{-}NH_2 + COCl_2 \xrightarrow{-HCL} \underset{(\mathbf{1})}{R{-}NHCOCl} \xrightarrow{\Delta} R{-}N{=}C{=}O + HCl} \quad (1)$$

first, leading to carbamoylchloride (**1**), is exothermic, while the hydrogen chloride (HCl) abstraction to form the isocyanates requires heat. The excess of phosgene prevents or suppresses the formation of side products, such as ureas. The industrial manufacture of isocyanates uses continuous processes in which streams of amine and phosgene, optionally both in solvents, are mixed and the resultant reaction mixtures heated, while the by-product hydrogen chloride is vented off. Products are typically purified by (vacuum) distillation. Other methods of preparation, such as the reaction of salts of isocyanic acid with alkyl halides or the thermal or photolytic degradation of carbonylazides, are less common and are used only in special cases.

The direct conversion of aromatic amines or nitro compounds with carbon monoxide (CO) in the presence of platinum metal catalysts [reaction (2)]

$$\mathrm{R{-}NO_2 + 3CO \longrightarrow R{-}N{=}C{=}O + 2CO_2} \quad (2)$$

has been explored as a viable alternative for the industrial manufacture of isocyanate in place of the phosgenation technology, which is less desirable in terms of the effect on the environment. However, technical problems have so far prevented the commercialization of any of the various proposed routes.

Characterization. The heterocumulene system N═C═O absorbs strongly in the infrared region at approximately 2250–2275 cm^{-1}; this band is very characteristic and is widely used for the spectroscopic identification of isocyanates. Isocyanates react readily with alcohols to form urethanes (also known as carbamates), which are used for their characterization. *See* INFRARED SPECTROSCOPY.

Reactions. Isocyanates are very reactive compounds. They have been used extensively as building blocks of nitrogen heterocycles, especially in the synthesis of fused-ring systems. Typically, the C═N double bond of the isocyanate group reacts readily with active hydrogen-containing substrates, such as amines, alcohols, or water. Alcohols react with isocyanates to form urethanes [reaction (3)] while primary amines give *N*,*N′*-disubstituted ureas. In the latter reaction, water adds to give initially *N*-substituted carbamic acids, which undergo decarboxylation at room temperature to re-form amine and carbon dioxide [reaction(4)]. Most often, excess isocyanate consumes the liberated amine to yield symmetric disubstituted ureas [reaction (5)].

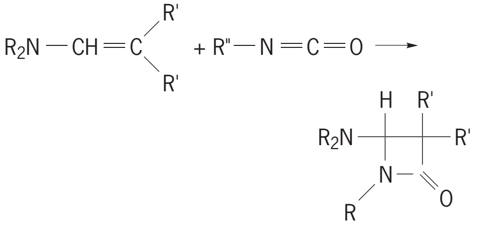
$$\text{R—N═C═O} + \text{R}'\text{OH} \longrightarrow \text{R—NACOOR}' \quad (3)$$

$$(\text{R—N═C═O} + \text{H}_2\text{O} \longrightarrow [\text{R—NHCOOH}] \longrightarrow \text{R—NH}_2 + \text{CO}_2 \quad (4)$$

$$\text{RNH}_2 + \text{R—N═C═O} \longrightarrow \text{RNH—CONHR} \quad (5)$$

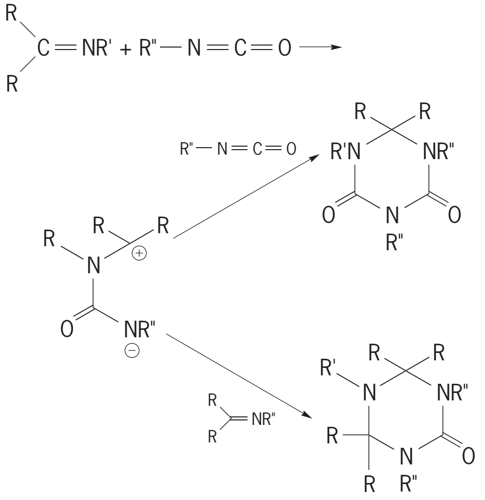
See AMINE; UREA.

Isocyanates readily undergo cycloaddition reactions with other compounds having activated double bonds, such as C═C, C═N, or C═O. The simplest such adducts are 2 + 2 cycloadducts, in which four-membered ring heterocycles are formed from an enamine [reaction (6)], a compound in which amino groups are directly linked to carbon-to-carbon double bonds.



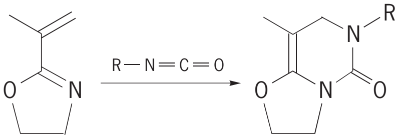
(6)

More common, however, are adducts formed from one double-bond component and two molecules of isocyanate, or vice versa. The reactions have been shown to proceed stepwise via dipolar intermediates. An example is the formation of a hexahydro-1,3,5-triazine derivative from an imine and isocyanate [reaction (7)]. The 1,4-dipolar intermediate can be intercepted by either a second molecule of isocyanate or an imine.

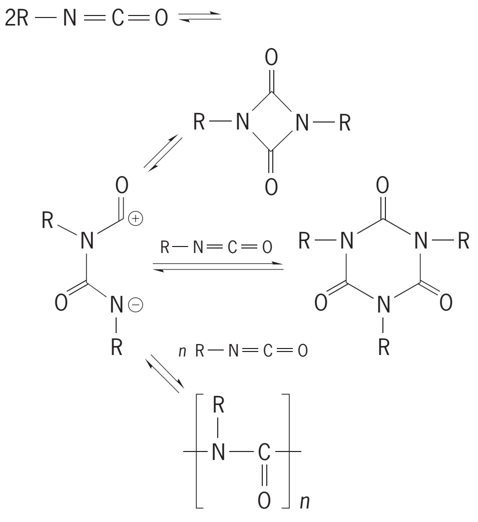


(7)

Systems containing conjugated double bonds have been shown to react with (predominantly) aryl and sulfonyl isocyanates across both double bonds in a hetero-diene synthesis; an example is shown in reaction (8), the formation of an oxazolopyrimidine from 2-vinyloxazole. Isocyanates are also known to react with each other to form low-molecular-weight oligomers or polymers under certain catalyzed conditions. Most commonly they form dimers, 1,3-diazetidinediones or trimers with hexahydro-1,3,5-triazinetrione structure, also known as isocyanurates [reaction (9)]. These cyclizations have also been shown to involve 1,4-dipolar intermediates; they are thermally reversible, especially in the case of the cyclodimers which dissociate back to starting isocyanates on heating to approximately 180–200°C



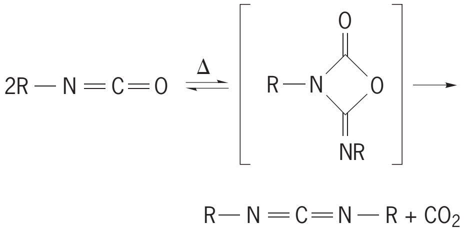
(8)



(9)

(350–390°F). An extension of this reaction type is the catalyzed polymerization of alkyl and aryl isocyanates to polyureas or 1-nylons at low temperature. These homopolymers, which are stabilized by addition of base as chain terminators, are thermolabile and usually melt with decomposition at 180–200°C (350–390°F); they have not found industrial application. *See* POLYMERIZATION.

The thermal generation of *N*,*N*′-disubstituted carbodiimides from isocyanates in the absence of catalysts is likely to involve a labile asymmetric isocyanate dimer in which the C═O group of the heterocumulene is involved in the cycloadduct formation [reaction (10)]. Other heterocumulenes, such as carbodiimides (RN═C═NR), isothiocyanates (RN═C═S), and ketenes (R_2C═C═O), are also known to undergo cycloadditions with isocyanates.

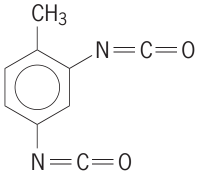


(10)

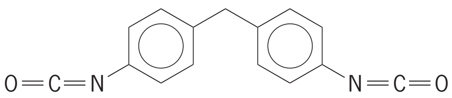
Isocyanates react with organometallic and phosphorus compounds primarily by inserting into the metal-X bond (X = OR, H, halogen, NR_2, and so forth). Several of these additions, especially those involving certain tin (Sn) compounds with SnOR or SnOSn bonds, have been studied in detail, since they exhibit catalytic activity in many isocyanate reactions such as trimerization and urethanization. *See* ORGANOMETALLIC COMPOUND.

Isocyanate-based polymers. Polymers based on the reactions of isocyanates were first developed in 1937. Diisocyanates react with difunctional reagents, such as diols, to form addition polymers with a wide variety of properties. The flexibility in the choice of starting materials (diisocyanate, diol, diamine, diacid, and so forth) and consequently in the multitude of possible adducts make this product group unique in the field of polymeric materials.

Two aromatic diisocyantes, tolylene diisocyanate (TDI; **2**) and di(4-isocyanatophenyl)methane (MDI; **3**) have become the major starting materials for a family of polymeric products, such as flexible and rigid polyurethane foams used in construction and appliance insulation, automotive seating, and furniture. Elastomers based on MDI, polyols, and polyamines are widely used in the automotive industry, where reaction injection molding technology is used for the manufacture of exterior parts such as body panels and bumpers. *See* POLYOL.



(**2**)



(**3**)

Thermoplastic polyurethane elastomers (TPU) are used in the molding and extrusion of many industrial and consumer products with superior abrasion resistance and toughness. *See* POLYURETHANE RESINS.

The trimerization to polyisocyanurates and the formation of polyamides from dicarboxylic acids have been utilized to synthesize polymers with excellent thermal properties. Aliphatic diisocyanates, notably 1,6-diisocyanatohexane (HDI), fully hydrogenated MDI (H_{12}MDI), and isophorone diisocyanate (IPDI), have become building blocks for color-stable polyurethane coatings and elastomers with high abrasion resistance. *See* POLYAMIDE RESINS; POLYMER.

Toxicology. Most low-molecular-weight isocyanates are strong lacrimators, and exposure to vapors should be avoided. Inhalation of vapors can cause pulmonary edema. Isocyanates are also known to be strong skin irritants that can cause allergic eczema and bronchial asthma, and they have been known to produce severe skin and lung hypersensitivity. Most isocyanates have a low level of oral toxicity. Amine precursors of several industrial diisocyanates are believed to be carcinogenic. Exposure limits have been set for isocyanate vapors in industrial environments. Reinhard H. Richter

Bibliography. V. I. Gorbatenko and L. I. Samarai, Synthesis and reactions of haloalkyl isocyanates, *Synthesis*, 1980:85–110, 1980; G. Oertel, *Polyurethane Handbook*, 2d ed., 1993; S. Patai (ed.), *The Chemistry of Cyanates and Their Thioderivatives*, 1977; A. A. R. Sayigh, H. Ulrich, and W. J. Farrissey, Jr., Diisocyanates in J. K. Stille (ed.), *Condensation Polymers*, pp. 369–476, 1972; H. Ulrich, Synthesis of polymers from isocyanates in polar solvents, *J. Polym. Sci.*: *Macromol. Rev.*, 11:93–133, 1976.

Isoelectric point

The pH of the acidity or alkalinity of the dispersion medium of a colloidal suspension or an ampholyte at which the solute does not move in an electrophoretic field. The term isoelectric point is abbreviated pI.

Ampholytes. These molecules carry acid as well as basic functional groups. When dissolved in a suitable medium, ampholytes may acquire positive and negative charges by dissociation or by accepting or losing protons, thereby becoming bipolar ions (zwitterions). Ampholytes may be as small as glycine and carry just one chargeable group each; or as large as polyampholytes (polyelectrolytes that carry positive charges, negative charges, or both). They may possess molecular weights in the hundreds of thousands like proteins or in the millions like nucleic acids, and carry many hundreds of chargeable groups. *See* ION; NUCLEIC ACID; PH; PROTEIN.

Determination. An example of establishing the isoelectric point is shown by the course of the pH changes during the titration of alanine [$NH_2CH(CH_3)COOH$], a 1:1 ampholyte, meaning a molecule that carries one positively and one negatively ionizable group. Starting from acid solution (**Fig. 1**), relatively small pH changes (with alkali as the titrant) are observed between pH 2 and 3 (acidic), and again between pH 9.5 and 10.5 (alkaline), caused by the buffering capacity of the carboxyl (—COOH) and amine (—NH_2) groups as weak electrolytes. The pH for $^1/_2$ equivalence corresponds to the pK of the acid function (a value related to the equilibrium constant), where one-half of the alanine molecules still carry only a positive charge ($-NH_3^+$), while the other half are also negatively charged ($-COO^-$). Thus alanine exists in the form of zwitterions. *See* PK.

In the case of alanine, the titration curve is repeated with close symmetry on its alkaline side. After adding 1.5 equivalents of sodium hydroxide (NaOH), the pH is that of the pK of the basic groups, which is also the range of maximum buffering capacity for alkaline pHs. At this point, one-half of the alanine molecules are still in the zwitterionic state, and one-half are entirely negatively charged.

Halfway between the acid and alkaline titration ranges, after adding 1 equivalent of NaOH and thus creating acid-base neutrality, all alanine molecules are in the zwitterionic, net-zero-charge state, and the $pH \equiv pI$. The value of pI in this example can be calculated as the average of the two [basic (b) and acidic (a)] pK values by the equation

$$pI = \frac{pK_a + pK_b}{2} = 6.1$$

The net charges of alanine vary in an easily recognizable way from the far acid to far alkaline pHs, because the 2 pK values of the two chargeable groups lie far apart. The same is true for the titration curve of the amino acid lysine, $NH_2CH[(CH_2)_4NH_2]COOH$, a 2:1ampholyte, meaning a molecule that carries three ionizable (chargeable) groups, two of which are the same and one of opposite sign (**Fig. 2**). The inflection points mark the pK's. The pH at which the curve of overall charge crosses from positive to negative values marks the isoelectric point at pH = 9.74 = pI. Here, lysine is in the purely zwitterionic state.

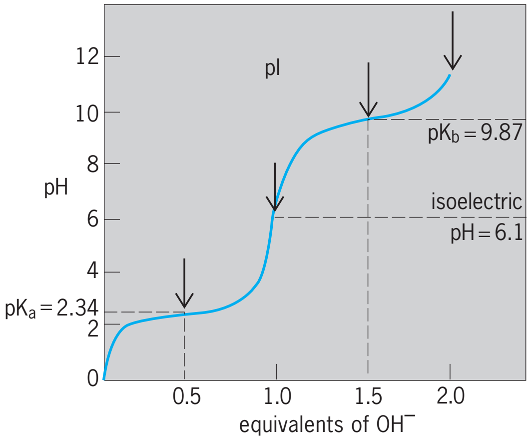


Fig. 1. Titration of alanine with sodium hydroxide (NaOH), showing the course of pH with added fractional equivalents. The four arrows show, from left to right, the pK_a [$^1/_2$ cations: $NH_3^+(CH_3)COOH$, and $^1/_2$ zwitterions: $NH_3^+(CH_3)CHCOO^-$], the pI (all zwitterions), the pK_b ($^1/_2$ zwitterions, $^1/_2$ anions), and the end of titration, when all alanine molecules are in the anionic form: $NH_2(CH_3)CHOO^-$.

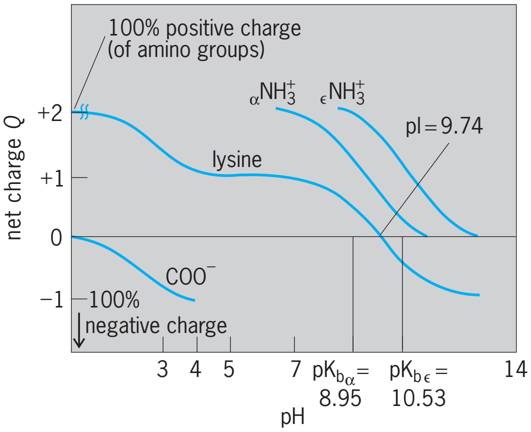


Fig. 2. Course of the charges carried by lysine ($pK_a = 2.18$) molecules in dilute aqueous solution with pH during a titration with alkali. Shown are the net charge, going from +2 at low pH to +1 between pH 4.5 and 6.5, the isoelectric point at pH 9.74, the net charge approaching −1 at about pH 13, and the courses of the individual charges on the three chargeable groups. The break in the curve indicates that in this range the course of the net charges (Q) is not shown completely, but eventually reaches a value of +2.

For molecules that carry four or more chargeable groups, that is, for polyelectrolytes, the courses of the overall titration curves may no longer reflect the individual dissociation steps clearly, as the dissociation areas usually overlap. The isoelectric point then becomes an isoelectric range, such as for pigskin (parent) gelatin, a protein that exhibits an electrically neutral isoelectric range from pH 7 to pH 9.

Since ampholytes in an electric field migrate according to their pI with a specific velocity to the cathode or anode, the blood proteins, for example, can be separated by the techniques of gel or capillary-zone electrophoresis. *See* ELECTROPHORESIS; TITRATION.

Isoelectric focusing. In the method of isoelectric focusing, a pH gradient within a chromatographic column parallels the potential gradient. The components (for example, proteins) of a mixture will each migrate in the electric field until they reach their respective isoelectric zone, at which point they will stop moving. Different components thus become neatly separated along the electrophoretic path.

Influence of change in reactivities. Since the pI depends on the pKs of the individual groups of an ampholyte, anything that affects the reactivities of the groups also affects the pI. This is the case when a variety of cosolutes, especially electrolytes, are present. An example is gelatin, whose midpoint of the isoelectric range will move from 8 to 6 if the concentration of codissolved sodium chloride (NaCl) changes from 10^{-3} to 10^{-1} *N*. Similarly, the solubility of an ampholyte, which depends on the solvation of its charge-carrying groups, varies with pH, with codissolved electrolyte, or with other cosolutes. At the isoelectric point, many ampholytes show a maximum of solution turbidity, some form of precipitation, or a larger friction factor and a maximum or minimum of viscosity. *See* CONCENTRATION SCALES; ELECTROLYTE.

Ion-exchange chromatography. The important separation technique of ion-exchange chromatography is based on the selective adsorption of ampholytes on the resins with which the column is filled, at a given pH. For example, the larger the net positive charge of an ampholyte, the more strongly will it be bound to a negative ion-exchange resin and the slower will it move through the column. By rinsing with solutions of gradually increasing pH, the ampholytes of a mixture can be eluted and made to emerge separately from the column and be collected. Automated amino acid analyzers are built on this principle.

Significance. The notion that some ampholytes may pass with changing pH through a state of zero charge (zero zeta potential) on their way from the positively to the negatively charged state has become so useful for specifying and handling polyampholytes that it was extended to all kinds of colloids, and to solid surfaces that are chargeable in contact with aqueous solutions. Practically all metal oxides, hydroxides, or hydroxy-oxides become charged by the adsorption of hydrogen ions (H^+) or hydroxide ions (OH^-), while remaining neutral at a specific pH. Strictly speaking, the isoelectric point of electrophoretically moving entities is given by the pH at which the zeta potential at the shear plane of the moving particles becomes zero. The point of zero charge at the particle (solid or surface) is somewhat different but often is not distinguished from the isoelectric point. It is determined by solubility minima or, for solid surfaces, is found by the rate of slowest adsorption of colloids (for example, latexes) of well-defined charge. *See* AMINO ACIDS; COLLOID; ELECTROKINETIC PHENOMENA; ION EXCHANGE; ION-SELECTIVE MEMBRANES AND ELECTRODES. F. R. Eirich

Bibliography. P. C. Hiemenz, *Principles of Colloid and Surface Chemistry*, 3d ed., 1997; J. W. Jorgenson, New directions in electrophoretic methods, *Amer. Chem. Soc. Symp.*, 335:182–198, 1987; N. Kallay et al., Determination of the isoelectric points of several metals by an adhesive method, *J. Phys. Chem.*, 95:7028–7032, 1991; P. G. Rhigetti, *Isoelectric Focusing*, 1983; C. Tanford, *Physical Chemistry of Macromolecules*, 1961.

Isoelectronic sequence

A term used in spectroscopy to designate the set of spectra produced by different chemical elements ionized in such a way that their atoms or ions contain the same number of electrons. An atom or ion with the same number of electrons will generally have spectra with similar features, as the number of electrons determines most atomic properties.

The sequence in the **table** is an example. Since the neutral atoms of these elements each contain *Z* electrons, removal of one electron from scandium, two from titanium, and so forth, yields a series of ions all of which have 20 electrons. Their spectra are therefore qualitatively similar, but the spectral terms (energy levels) increase approximately in proportion to the square of the core charge, just as they depend on Z^2 in the one-electron sequence H, He^+, Li^{2+}, and so forth. As a result, the successive spectra shift progressively toward shorter wavelengths (higher energies), but maintain similar spectral features. Isoelectronic sequences are useful in predicting unknown spectra of ions belonging to a sequence in which other spectra are known. *See* ATOMIC STRUCTURE AND SPECTRA.

F. A. Jenkins; W. W. Watson; Alfred S. Schlachter

Example of isoelectronic sequence

Designation of spectrum	Emitting atom or ion	Atomic number, Z	Number of electrons
CaI	Ca	20	20
ScII	Sc^+	21	20
TiIII	Ti^{2+}	22	20
VIV	V^{3+}	23	20
CrV	Cr^{4+}	24	20
MnVI	Mn^{5+}	25	20

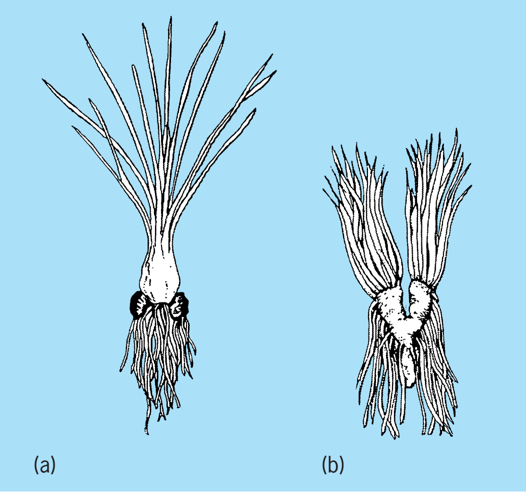
Isoetales

An order in the class Lycopsida that diverged from the Lepidodendrales in the Late Devonian. These two groups have several characters that are not found in other lycopsids, notably a centralized, shootlike rooting structure (rhizomorph) that allows finite growth, wood production, and tree-sized dimensions. Isoetaleans evolved from trees as an increasingly specialized and reduced lineage; all but the earliest are small-bodied shrubs and pseudoherbs. A reduced morphology characterizes the only living isoetalean genus, *Isoetes*. *See* LEPIDODENDRALES.

Isoetes (see **illus.**) is a globally distributed genus of approximately 150 species. Many are poorly differentiated, as their simple morphology lacks taxonomically useful characters and hence indicates a high probability of evolutionary convergence.

The slow-growing plants resemble a small grass or rush. The basal corm is a highly reduced, woody, bipolar rhizomorph, anchored in the substrate by rootlets and capped by a vestigial stem. It is surrounded by clasping quill-like leaves that function as sporophylls, developing large sporangia. As resources decline during the growing season, the sporangia that are formed on successive sporophylls switch from being “female” megasporangia to “male” microsporangia, and eventually to abortive sporangia. Megasporangia contain several hundred megaspores, whereas microsporangia contain numerous small microspores that germinate to release spermatozoids that are multiflagellate.

Isoetaleans are delimited by the bilateral symmetry and ontogeny of their rhizomorphs, relative to the



Morphology and anatomy of extant *Isoetes*. (*a*) Compact, cormose rhizomorph and unbranched stem of *I. melanopoda*. (*b*) More elongate, once-branched rhizomorph and stem of *I. andicola*. (*After W. N. Stewart and G. W. Rothwell, Paleobotany and the Evolution of Plants, 2d ed., Cambridge University Press, 1993*)

radial rhizomorph symmetry of the Lepidodendrales. The most likely origin of the Isoetales is from among the relatively primitive lepidodendraleans that possess bisporangiate cones resembling those of the Selaginellales. The group originated in the Late Devonian with poorly known small trees such as *Lepidosigillaria, Cyclostigma*, and later *Protostigmaria-Lepidodendropsis*. The Late Carboniferous *Sporangiostrobus* and *Chaloneria* were small trees and shrubs inhabiting coal swamps. The Early Triassic *Pleuromeia* and Early Cretaceous *Nathorstiana* (Pleuromeiales) resembled *Chaloneria* in growth habit but showed reductions of the rhizomorph. Plants resembling *Isoetes* occur in the Middle to Late Triassic periods, but the earliest definite *Isoetes* species are of Early Cretaceous age. *See* SELAGINELLALES.

Isoetaleans inhabited Paleozoic coal swamps (*Chaloneria*), Mesozoic coastal marshes (*Pleuromeia*), and Cenozoic river and lake systems (*Isoetes*). They typically formed a major ecological component of low-diversity communities that lacked large trees. *See* LYCOPHYTA; LYCOPSIDA.

Richard M. Bateman; William A. DiMichele

Bibliography. K. B. Pigg, Evolution of isoetalean lycopsids, *Ann. Mo. Bot. Gard.*, 79:589–612, 1992; J. E. Skog and C. R. Hill, The Mesozoic herbaceous lycopsids, *Ann. Mo. Bot. Gard.*, 79:648–675, 1992; W. C. Taylor and R. J. Hickey, Habitat, evolution, and speciation in *Isoetes, Ann. Mo. Bot. Gard.*, 79:613–622, 1992.

Isolaimida

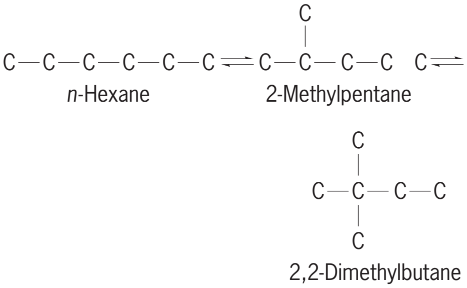
An order of nematodes comprising the single superfamily Isolaimoidea. The order consists of one family and one genus. They are rather large for free-living soil nematodes (0.1–0.2 in. or 3–6 mm) and are found in seldom-cultivated sandy soils. Some forms have anterior annulations, while others have posterior transverse rows of punctations. The diagnostic characteristics of this superfamily are the presence of six hollow tubes around the oral opening and two whorls of six circumoral sensilla. Amphids are apparently absent, though some authors speculate that their function is taken over by the dorsolateral papillae of the second whorl. The triradiate stoma is elongate and has thickened walls anteriorly. The esophagus is clavate. The female ovaries are opposed and outstretched, and may have flexures. The gubernaculum, supporting the thickened male spicules, has two dorsal apophyses, and male preanal supplementary organs are papilloid. Paired caudal papillae are large in males, small in females. *See* NEMATA (NEMATODA).

Armand R. Maggenti

Isomerization

Rearrangement of the atoms within hydrocarbon molecules. Isomerization processes of practical significance in petroleum chemistry are (1) migration of alkyl groups, (2) shift of a single-carbon bond in naphthenes, and (3) double-bond shift in olefins.

Migration of alkyl groups. An example of alkyl group migration (skeletal isomerization) is reaction (1). Isomerization to more highly branched configurations has commercial importance since it results in improvement in combustion quality in the automobile engine as measured by octane number and increased chemical reactivity because tertiary carbon atoms result. The unleaded, motor-method octane numbers of the hexane isomers shown in reaction (1) are 26.0, 73.5, and 93.4, respectively. Normal butane is converted to isobutane (which has a tertiary carbon atom) to attain chemical reactivity with olefins in alkylation reactions where *n*-butane is inert.



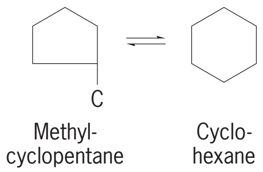
(1)

Isomerization of paraffins is a reversible first-order reaction limited by thermodynamic equilibrium which favors increased branching at lower temperatures. Undesirable cracking reactions leading to catalyst deactivation occur at higher temperatures. They are controlled by adding a cracking suppressor such as hydrogen.

Conversion of normal butane to isobutane is the major commercial use of isomerization. Usually, it is carried out in either liquid- or vapor-phase over aluminum chloride catalyst promoted with hydrogen chloride. In the vapor-phase process (250–300°F or 120–150°C), the aluminum chloride is often supported on bauxite. In the liquid-phase processes (180°F or 82°C), it is dissolved in molten antimony trichloride or used in the form of a liquid complex with hydrocarbon. A second type of catalyst for vapor-phase isomerization (300–850°F or 150–450°C) is a noble metal, usually platinum, supported on a carrier. This may be alumina with halide added to provide an acidic surface. All the processes are selective (95–98% to isobutane). Approximately 60% of the *n*-butane feed is converted per pass to isobutane in the liquid-phase process.

Isopentane, a high-octane component used in aviation gasoline, is made commercially by isomerization of *n*-pentane. Petroleum naphthas containing five- and six-carbon hydrocarbons also are isomerized commercially for improvement in motor-fuel octane numbers. Noble-metal catalyst is normally used with higher-molecular-weight feeds. Isomerization of paraffins above six carbon atoms is of less importance, since octane improvement is limited by predominance of monomethyl branching at equilibrium. Skeletal isomerization is an important secondary reaction in catalytic cracking and catalytic reforming. Aromatics and olefins undergo skeletal isomerization as do paraffins.

Single-carbon bond shift. This process, in the case of naphthenes, is illustrated by reaction (2). Cyclohexane and methylcyclohexane have been produced commercially by liquid-phase isomerization of the five-carbon ring isomers over aluminum chloride–hydrocarbon-complex catalyst promoted by hydrogen chloride. Conversion per pass is high, selectivity excellent, and reaction conditions mild (190°F or 88°C). Cyclohexane is a raw material for making nylon, and it may be dehydrogenated to benzene. Methylcyclohexane has been used to make synthetic nitration-grade toluene.



(2)

Shift of a double-bond. This process is usefully applied when a specific olefin is needed for chemical synthesis, as in reaction (3). Double-bond shift

$$\mathrm{C{-}C{-}C{=}C \rightleftharpoons C{-}C{=}C{-}C} \qquad (3)$$

occurs selectively over acidic catalysts at temperatures below 450°F (230°C). However, the proportion undergoing skeletal isomerization increases as temperature is increased until, at temperatures in the range of 600–950°F (300–510°C), equilibrium is approached at fairly high space velocities. Equilibrium favors movement of double bonds to the more stable internal positions (85.6% 2-butene at 400°F or 200°C), and octane improvement accompanies this shift; however, the increase of octane number normally is insufficient to justify the cost of processing thermally cracked gasolines solely for this purpose. This type of isomerization occurs as a secondary reaction in the catalytic cracking and catalytic polymerization processes, in part accounting for the high octane numbers of the gasolines. *See* AROMATIZATION; CRACKING; MOLECULAR ISOMERISM; PETROLEUM PROCESSING AND REFINING. George E. Liedholm

Isometric process

A constant-volume thermodynamic process in which the system is confined by mechanically rigid boundaries. No direct mechanical work can be done on the surroundings by a system with rigid boundaries; therefore the heat transferred into or out of the system equals the change of internal energy stored in the system. This change in the internal energy, in turn, is a function of the specific heat and the temperature change in the system as in Eq. (1), where Q_V

$$Q_V = U_2 - U_1 = \int_1^2 C_V dT \qquad (1)$$

is the heat transferred at constant volume, U is the internal energy, C_V is the heat capacity at constant volume, and T is the absolute temperature. If the process occurs reversibly (the system going through a continuous sequence of equilibrium states), Eq. (2)

$$Q_V = \int_1^2 T dS \qquad (2)$$

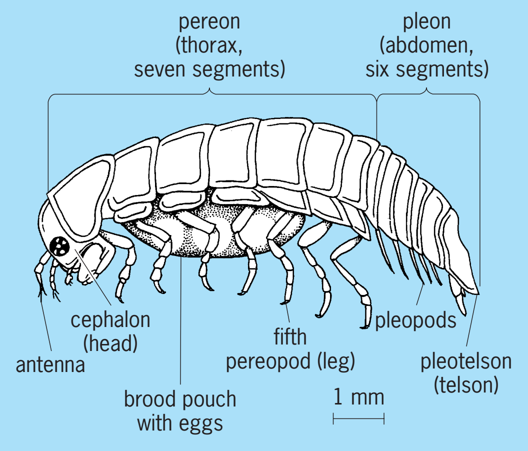
holds, where S is the entropy. There is an increase in both the temperature and the pressure of a constant volume of gas as heat is transferred into the system. For a comparison of the isometric process with other processes involving a gas *see* POLYTROPIC PROCESS Philip E. Bloomfield

Isopoda

An order of invertebrate crustaceans related to crabs, shrimps, and lobsters, all members of the class Malacostraca. The closest relatives of Isopoda are the amphipods, mysids (opposum shrimps), cumaceans, and tanaids, all of which are placed in the superorder Peracarida. Isopods are generally small but very common, highly diversified, and occurring in marine, freshwater, and terrestrial habitats. Sow bugs, pill bugs, and woodlice, as well as their marine relatives (such as gribbles and sea slaters), belong to this group. *See* CRUSTACEA; MALACOSTRACA; PERACARIDA.

Morphology. Isopods are characterized by the lack of a carapace (an outer case or covering made of chitin), having the first thoracic segment fused with the head, one pair of maxillipeds (one of the four sets of paired mouthparts used for feeding), and usually seven pairs of similar legs, although the first leg may be clawlike. Adults of the suborder Gnathiidea possess five pairs of legs.

As in most crustaceans, the body has three parts: the head (cephalon), thorax (pereon), and abdomen (pleon) [see **illustration**]. The cephalon bears sessile eyes and two pairs of antennae, the first of which is chemosensory and the second tactile in function. Each of the seven segments of the pereon



***Limnoria* (Flabellifera), female. (*After R. J. Menzies, The comparative biology of the wood-boring isopod crustacean Limnoria, Museum of Comparative Zoology, Harvard Coll. Bull., 112(5):363–388, 1954*)**

(pereonites) normally bears a pair of legs called pereopods. These are used for locomotion and grabbing prey. The abdomen typically consists of five segments or pleonites (which may be fused in some) plus a sixth segment called the pleotelson. Uropods are the paired appendages at the base of the pleotelson; these are used in orientation and swimming. Each of the pleonites has a ventral pair of foliaceous appendages (leaflike pleopods) which are used for swimming and respiration.

Ecology. At present there are more than 10,400 described species of isopods. Of those, about 4200 species are terrestrial, and these include the familiar sow bugs, pill bugs, roly-polys, and woodlice. They may be found in any habitat on land, including forests, caves, deserts, city sidewalks, and greenhouses.

Terrestrial isopods feed on algae, moss, bark, and decaying plant and animal matter. Like insects, they play a role in the breakdown and recycling of organic nutrients. Cellulose digestion results from the presence of bacteria in the digestive system.

In the continental United States and Canada, there are about 120 known species of terrestrial isopods. The most common are widely distributed and have probably been introduced from Europe. A majority of the most abundant species in North America are concentrated in the northeast and northwest, possibly a result of the transport of their ancestors into harbors by human immigrants.

Marine isopods, numbering more than 5600 species, are found at all latitudes and in all marine habitats, from the intertidal zone to the deep sea, including coral reefs, sponges, algae, and seagrass beds. They range in size from less than a millimeter (<0.04 in.) [one of the parasitic forms] to 1.5 mm (0.06 in.) [*Microcharon*, found living between sand grains] to 44 cm (17.3 in.) [*Bathynomus* in the depths of the Caribbean Sea]. A few, such as *Idothea*, are cosmopolitan (having worldwide distribution), perhaps being distributed by floating on rafts of seaweed. Others travel in ballast water in oceangoing ships from place to place.

Because they have limited swimming ability and move chiefly by crawling on the bottom, most isopods have relatively small ranges. They recycle nutrients by eating detritus, both plant and animal material, and become part of the food chain in the sea as they are eaten by larger animals, particularly fish. Marine isopods may be omnivores, herbivores, or predators; some specialize as parasites, living off fluids and tissues of their hosts, mainly fish and other crustaceans. They fulfill many of the same ecological niches as insects.

The isopod species found in freshwater total about 600. Their habitats include rivers, subterranean streams, caverns, wells, swamps, and hot springs. They tend to have very restricted habitats and thus are the most likely isopods to be designated as threatened or endangered species.

Reproductive biology. As is typical of the peracarids, isopods develop inside the female brood pouch (marsupium) located on the underside of the abdomen. When the young emerge, they typically lack the seventh pair of legs and are known as manca but otherwise look like adults (except in Gnathiidea). Some evidence of parental care has been recorded in the family Sphaeromatidae (suborder Flabellifera); and in *Iais* (suborder Asellota) adults are known to carry the manca about by clasping it to their ventral abdomen. Mate guarding has also been observed. Examples of sequential hermaphroditism or sex change have been found in several groups, both male-to-female and female-to-male. However, sexes are usually separate and can be virtually identical or widely divergent in morphology, depending on the suborder or family.

Evolution. Isopods date from the Paleozoic Era, when malacostracans as a group underwent a rapid evolution. The earliest fossil records of isopods contain members of the present-day Phreatoicidea, dating 325 million years ago, representing a very early offshoot from the original common ancestor. These animals, currently found only in freshwater habitats, are only slightly changed from their fossil ancestors and may be considered as living fossils (living species belonging to an ancient stock otherwise known only as fossils). Other isopod groups probably evolved as benthic (bottom) dwellers in shallow marine environments such as estuaries and embayments before radiating into the deep sea, and they also conquered land as the only truly terrestrial crustaceans (suborder Oniscidea). *Cyclosphaeroma* from the Jurassic, discovered in England, is related to members of the contemporary family Sphaeromatidae.

Classification. The classification system currently in wide use recognizes 10 suborders, listed below.

Anthuridea are generally elongate and cylindrical, living in soft bottoms of estuaries and oceans, especially in tropical reef environments.

Asellota are very diverse, are generally herbivores, and live in freshwater as well as marine habitats. They are particularly diverse in the deep sea. One pair of pleopods forms a cover over the rest of the pleopods.

Calabozoida have been found thus far only in freshwater wells in Venezuela. They have unique body segmentation and pleopod structure.

Epicaridea are exclusively external parasites on other crustaceans and may be highly modified. The male is much smaller than the female, whose body may be very distorted.

Flabellifera inhabit both fresh and salt water; they can be free-living, predaceous, commensal, or parasitic. A common characteristic is a tail fan comprising the pleotelson and posterior appendages (uropods).

Gnathiidea can be found only in marine environments but at all latitudes. The male is known for large frontal mandibles. The juveniles are parasitic on fish, while the adults probably do not feed.

Microcerberidea may be found in marine, brackish, or freshwater in sand or sediment. They are less than 2 mm (0.08 in.) in length.

Oniscidea, the sow bugs or pill bugs, have successfully conquered almost all terrestrial habitats, as their pleopods are adapted for breathing air.

Phreatoicidea, relict (being a persistent, isolated remnant of a once-abundant species) crustaceans, are found thus far only in Australia, Tasmania, and South Africa. They are laterally compressed with long antennae and long uropods.

Valvifera have uropods which are flexed beneath the abdomen and cover the pleopods. Many live on seaweed or seagrass, from the tropics to the polar regions.

Several authors (A. Brandt and G. C. B. Poore) have suggested that the Flabellifera is an artificial grouping, so further reorganization of this system is certain. These authors have also erected two new suborders, Tainisopodidea, including members of the family Tainisopidae, found only in ground waters in western Australia, and Phoratopidea, comprising a single species found only in Australia. Marilyn Schotte

Bibliography. A. Brandt and G. C. B. Poore, Higher classification of the flabelliferan and related Isopoda based on reappraisal of relationships, *Invert. Systemat.*, 17(6):893–923, 2003; G. Brusca and R. C. Brusca, *Invertebrates*, Sinauer Associates, Sunderland, 1990; B. Kensley and M. Schotte, *Guide to the Marine Isopod Crustaceans of the Caribbean*, Smithsonian Institution Press, Washington, DC, 1989; B. Kensley, M. Schotte, and S. Schilling, *World List of Marine, Freshwater and Terrestrial Crustacea Isopoda*, www.nmnh.si.edu/iz/isopod; M. R. Warburg, *Evolutionary Biology of Land Isopods*, Springer-Verlag, Berlin, 1993.

Isoprostanes

A class of natural products discovered relatively recently. The isoprostanes (iPs) are isomeric with another class of natural products, the prostaglandins. The prostaglandins are the result of enzymatic oxygenation of polyunsaturated fatty acids (PUFAs), in particular arachidonic acid (AA). In contrast, the iPs are formed in vivo by a nonenzymatic, free-radical oxygenation of arachidonic acid. This important distinction in the mode of formation of the iPs is responsible for a more complex mixture of iPs being generated in vivo. For example, whereas the endoperoxide prostaglandin G_2 (PGG_2) is formed by the cyclooxygenase enzymes (COX1 and COX2), four classes of iPs are formed as a result of the free-radical oxygenation of arachidonic acid (**Fig. 1**), with each class containing 16 iPs for a total of 64 individual iP molecules. The discovery of iPs is important for two reasons: (1) group III iPs are incidental ligands for the prostaglandin receptors, hence they possess biological activity; (2) iPs are the product of oxidative stress. Their measurements have been shown to be a predictor of the onset and severity of inflammatory diseases such as Alzheimer's disease and atherosclerosis. *See* ALZHEIMER'S DISEASE; CIRCULATION DISORDERS; EICOSANOIDS; FREE RADICAL.

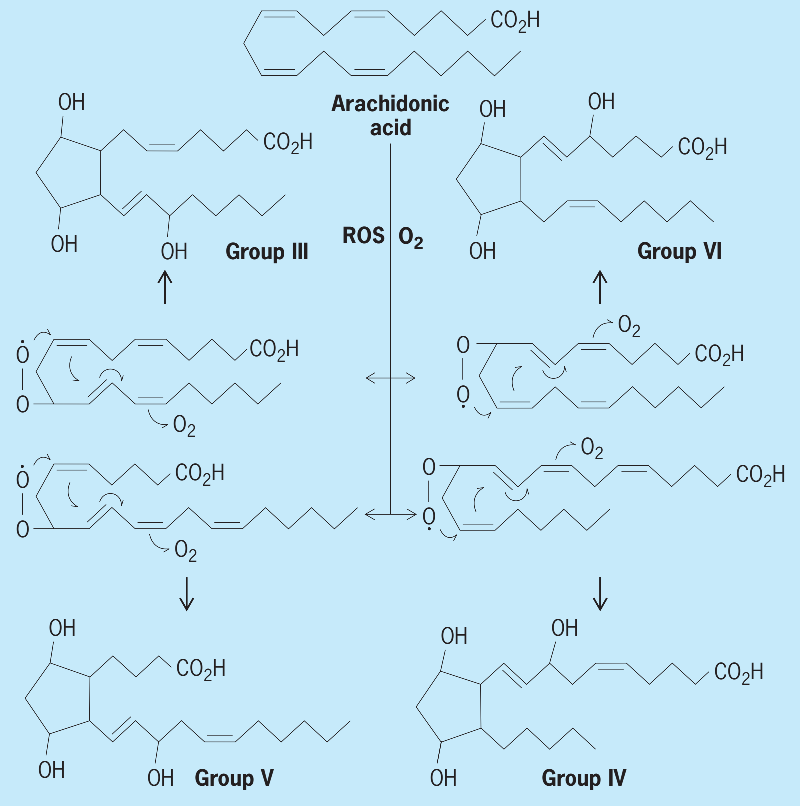


Fig. 1. Endoperoxide mechanism for the formation of isoprostanes. ROS = reactive oxygen species.

Mechanism for formation. Two mechanisms for the formation of iPs have been proposed based on in-depth regiochemical and stereochemical analysis of the peroxidation process. These are the endoperoxide mechanism (Fig. 1) and the dioxetane mechanism. Recently, evidence has been reported showing that the dioxetane mechanism may not be operative.

A more detailed mechanistic view for the formation of group VI iPs by the endoperoxide mechanism is presented in **Fig. 2**. The four iPs shown in the box have been prepared by total synthesis and used in discovery and identification of the same products in human urine (Fig. 4).

Nomenclature. A comprehensive nomenclature for iPs has been introduced utilizing the symbols in the prostaglandin nomenclature. The new classification

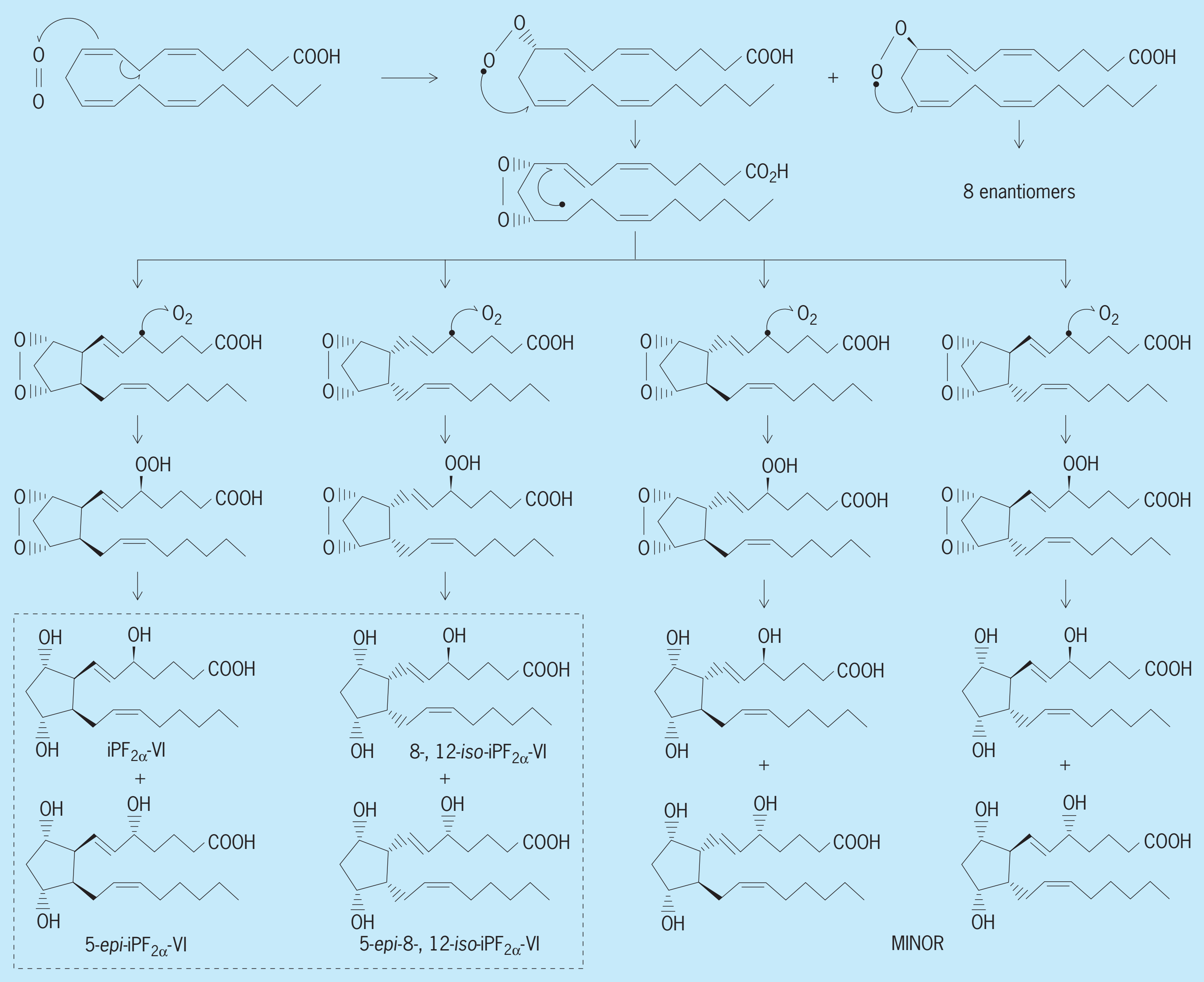


Fig. 2. Mechanism for the formation of group VI isoprostanes. There are 16 products, but for clarity 8 of the enantiomers are not shown.

(roman numerals) reflects the use of the *ω* end of the molecule instead of the COOH in the double-bond counting system—the same way that double bonds in polyunsaturated fatty acids are referred to as *ω*-3, *ω*-6, etc. Another nomenclature is also being used for iPs based on the location of the hydroxyl group in the side chain, the carboxyl end being C-1.

Measurement. Two techniques for measuring iPs are discussed below.

Gas chromatography/mass spectrometry (GC/MS) methods have been developed to quantitate individual iPs, particularly group VI iPs, which appear to be most abundant in urine. Group VI was chosen as the main initial focus because the position of the OH on carbon-5 of the upper side chain allows a lactone to form, which can easily be separated from the rest of the iPs, allowing for a more accurate and reliable determination in biological fluids (**Fig. 3**).

The GC/MS method outlined in Fig. 3 has enabled measurement of in vitro and in vivo elevated levels of $iPF_{2\alpha}$-VI in a number of clinical settings, such as ischemia-reperfusion syndromes, hypercholesterolemia, Alzheimer's disease, and alcohol-induced liver disease. Elevated levels of $iPF_{2\alpha}$-VI have been measured in oxidized low-density lipoprotein (LDL) and in aortic atherosclerotic lesions. In all cases, the discovery and identification of iPs of different groups, including those shown in Figs. 2 and 4, have been made possible by the initial availability of authentic material prepared by total synthesis. *See* GAS CHROMATOGRAPHY; MASS CHROMATOGRAPHY.

Liquid chromatography tandem mass spectrometry (LC/MS/MS) is another promising method for the identification, separation, and quantitation of iPs. Neither derivatization nor prior purification is required. Arachidonic acid–derived iPs readily generate abundant molecular ions (*m/z* 353) under ESI (electrospray ionization) conditions in the negative-ion mode, which fragments further to ions, some of which are group-specific (*m/z* 115, 193, 127, 151). **Figure 4** shows the chromatogram of the four groups of arachidonic acid–derived iPs from human

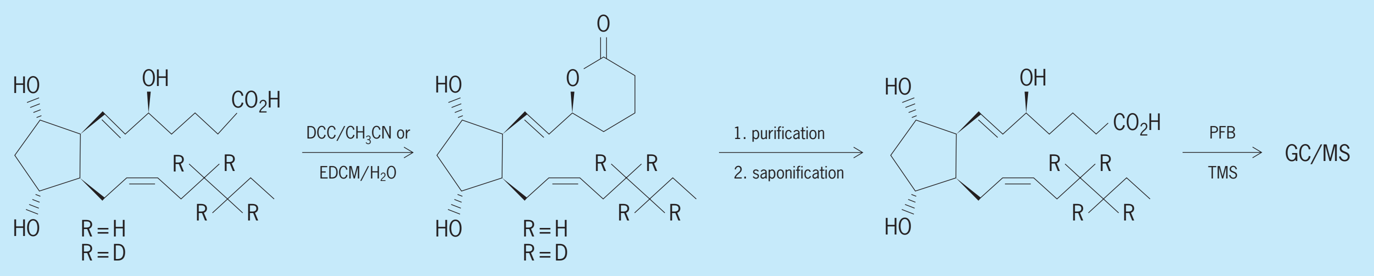


Fig. 3. GC/MS method measurement of $iPF_{2\alpha}$-VI. DCC = dicyclohexylcarbodiimide; EDCM = 1-ethyl-3(3-dimethylaminopropyl)carbodiimide methiodide; PFB = pentafluorobenzyl; TMS = trimethylsilyl; GC/MS = gas chromatography/mass spectrometry; H = hydrogen; D = deuterium.

urine. The chemical structure in the chromatogram identifies the individual iPs. The four most abundant iPs of urinary group VI have the cis configuration of the side chain. *See* LIQUID CHROMATOGRAPHY.

Other novel iPs and isofurans. Reports have shown that oxidation of eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA) result in the formation of F_3- and F_4-iPs. Since DHA is extremely rich in the brain, F_4-iPs, collectively called neuroprostanes, could be of great importance in assessing the severity of disease states such as Alzheimer's disease. Chemical syntheses of $iPF_{4\alpha}$-VI and $iPF_{4\alpha}$-III have been reported.

Recently, another series of isomeric compounds called isofurans has been discovered. These compounds, which contain a substituted tetrahydrofuran (THF) ring, are present at detectable levels in normal tissues and biological fluids. Increased amounts of these compounds have been observed in animal models exposed to high oxidant stress.

Biological implications. The importance of free-radical-generated iPs encompasses three general areas.

Damage to cell membranes. One consequence of oxidant injury is alteration in the physicochemical properties and function of cellular membranes. This has been attributed to an accumulation of products of lipid peroxidation in lipid bilayers. Phospholipids are essential constituents of the cell membrane. Radical oxygenation of the polyunsaturated fatty acids, located at the *sn*-2 position of the phospholipids, is bound to create disturbances in the cell membrane. The tightness of the fatty acid portion of the phospholipids is what causes the cell membrane to be hydrophobic and guarantees the integrity of the cell membrane. Polyhydroxylation of the unsaturated fatty acid to yield iPs, which contain three OH groups, is bound to disrupt the tightness of the membrane and have profound effects on the fluidity and integrity of the cellular membrane. For example, a leak could develop in the membrane due to the repulsion by the newly created OH groups and the adjacent lipophilic fatty acids, leading to cell death. *See* CELL MEMBRANES.

Biological activity. Although the mechanism by which iPs are released from the cell membrane remains to be elucidated, phospholipid-containing iPs are subject to cleavage by phospholipases. These iPs are released in the bloodstream, circulate through the liver and kidney, and eventually are excreted in urine, although their metabolic disposition in vivo remains largely unknown. Several iPs have been shown to have biological effects in vitro. For instance, $iPF_{2\alpha}$-III is a vasoconstrictor when infused into the rat renal artery and the rabbit pulmonary artery. In both cases the effects were prevented by pharmacological blockage of thromboxane receptors. Similarly, $iPF_{2\alpha}$-III modulates platelet function. Although it does not cause irreversible aggregation, it facilitates the ability of subthreshold concentrations of conventional agonists to induce this response. These effects require micromolar concentrations of $iPF_{2\alpha}$-III. Interestingly, much lower concentrations of this compound have been reported to modulate platelet adhesion.

Much less information is available about the potential biological activity of other iPs. For example, $iPE_{2\alpha}$-III behaves much like $iPF_{2\alpha}$-III on platelet aggregation. In contrast to these compounds, which appear to act as incidental ligands via the thromboxane receptor, 8,12-*iso*-$iPF_{2\alpha}$-III induces proliferative responses in NIH 3T3 cells via the $PGF_{2\alpha}$ receptor. This compound activates this prostanoid receptor selectively and in a saturable manner, but at concentrations roughly one order of magnitude greater than its natural ligand.

Given the difficulty of interpreting the biological potency of individual iPs when multiple species are likely to be formed coincidentally in vivo, it is premature to decide whether the biological properties of these compounds might be relevant under conditions of oxidant stress in vivo. Nonetheless, the possibility of coordinate cellular activation by multiple species, their delivery to cell membranes in highly concentrated form, and the superiority of thromboxane antagonism to cyclooxygenase inhibition in experimental settings suggest a potential role for these compounds as mediators as well as indices of oxidant stress.

Indices of oxidant stress. Perhaps the most interesting aspect of iPs research at present is the possibility that measurement of these compounds in biological fluids might afford a quantitative index of lipid peroxidation in humans. Increased plasma and urinary concentrations of $iPF_{2\alpha}$-VI and $iPF_{2\alpha}$-III have been reported in several syndromes thought to be associated with excessive generation of free radicals. These include coronary reperfusion after a period of ischemia, poisoning with paraquat and paracetemol,

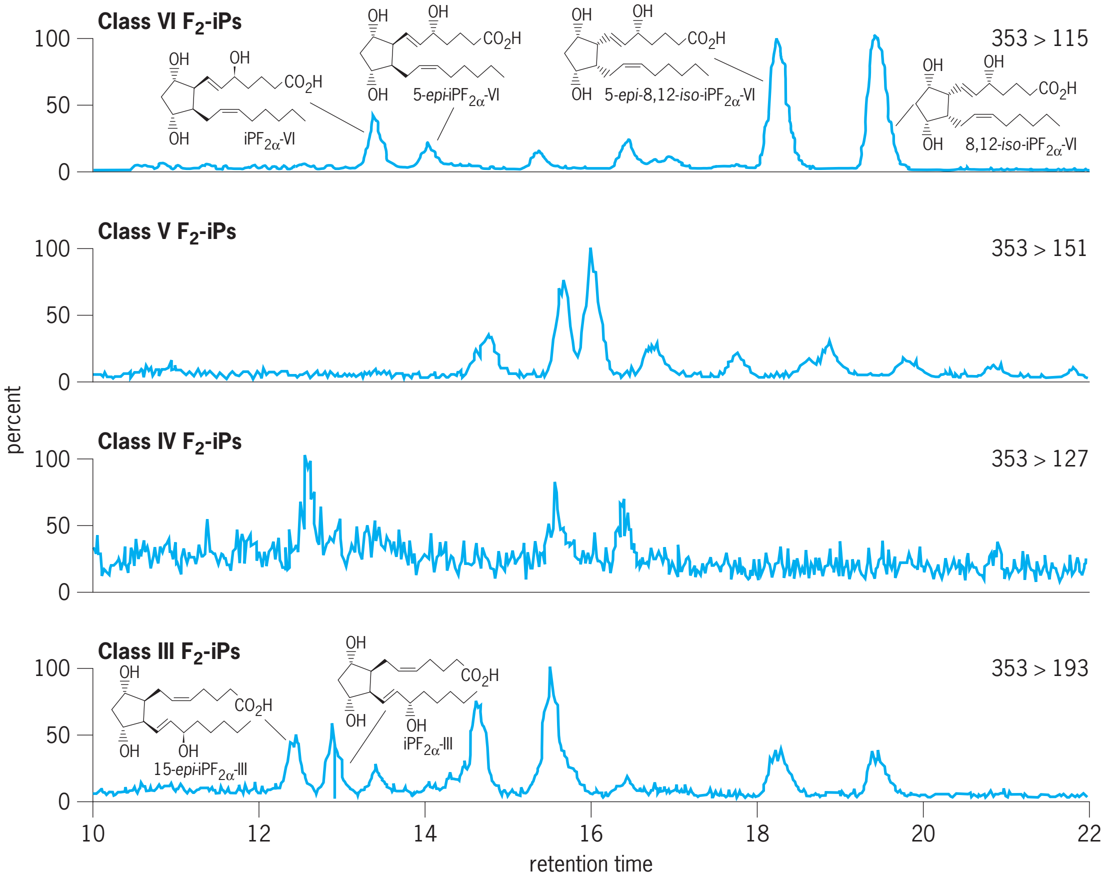


Fig. 4. LC/MS/MS of urinary isoprostanes. Selected peaks were identified by comparison with synthetic standards.

cigarette smoking-related illnesses, and alcohol-induced liver disease. Isoprostanes, which are increased in low density lipoprotein when it is oxidized in vitro, have been found to be increased in the urine of asymptomatic patients with hypercholesterolemia and are present in human atherosclerotic plaque. These observations, and their measurement in urine models of atherosclerosis, raise the likelihood that they may be useful in elucidating the role of LDL and cellular protein oxidation in atherogenesis. In perhaps the most striking evidence to date of the free-radical involvement in atherosclerosis, J. Rokach and G. A. FitzGerald have shown that suppression of elevated iP generation (iPF$_{2\alpha}$-VI) in vivo reduces the progress of atherosclerosis in apolipoprotein E–deficient mice. Significantly, they have also shown in this study that the level of cholesterol remains high and unchanged. It suggests that in cholesteryl linoleate, the entity present in LDL, the cholesterol moiety is the carrier of the linoleic acid which is responsible for the oxidative stress and the subsequent cell damage. Presumably the cholesterol part is left intact. Recently, the most abundant iP in urine, 8,12-*iso*-iPF$_{2\alpha}$-VI (Fig. 4), has been used as an index of oxidant stress and severity of Alzheimer's disease. *See* CHOLESTEROL.

Garret A. FitzGerald; Joshua Rokach

Bibliography. M. Adiyaman et al., Total synthesis of a novel isoprostane iPF$_{2\alpha}$-I and its identification in biological fluids, *Tetrahed. Lett.*, 37:4849–4852, 1996; J. P. Fessel et al., Discovery of lipid peroxidation products formed in vivo with a substituted tetrahydrofuran ring (isofurans) that are favored by increased oxygen tension, *Proc. Nat. Acad. Sci. USA*, 99:16713–16718, 2002; S. Kim et al., The first total synthesis of iPF$_{4\alpha}$-VI and its deuterated analog, *Tetrahed. Lett.*, 43:2801–2805, 2002; H. Li et al., Quantitative high performance liquid chromatography/tandem mass spectrometric analysis of the four classes of F2-isoprostanes in human urine, *Proc. Nat. Acad. Sci. USA*, 96:13381–13386, 1999; J. D. Morrow et al., A series of prostaglandin F2-like compounds are produced *in vivo* in humans by a non-cyclooxygenase, free radical-catalyzed mechanism, *Proc. Nat. Acad. Sci. USA*, 87:9383–9387, 1990; D. Pratico et al., Vitamin E suppresses isoprostance generation in vivo and reduces atherosclerosis in ApoE-deficient mice, *Nature Med.*, 4:1189–1192, 1998; L. G. Quan and J. K. Cha, Preparation of isoprostanes and neuroprostanes, *J. Amer. Chem. Soc.*, 124:12424–12425, 2002; J. Rokach et al., Nomenclature of isoprostanes: A proposal, *Prostaglandins*, 54:853–873, 1997; D. F. Taber, J. D. Morrow, and L. J. Roberts II, A nomenclature system for the isoprostanes, *Prostaglandins*, 53:63–67, 1997; E. A. Weinstein et al., Prothrombinase acceleration by oxidatively damaged phospholipids, *J. Biol. Chem.*, 275:22925–22930, 2000.

Isoptera

An order of the Insecta, commonly called termites, with the general characteristics and stages of other exopterygote insects. They are closely related to the Blattodea. Indeed, the only wood-eating cockroach, *Cryptocercus*, has a similar collection of protozoa in its hindgut to that of the primitive termite *Zootermopsis*. Similarly, the Australian *Mastotermes*, though socially complex, is considered the most primitive extant termite by reason of its having an anal lobe in the adult wing; this character is found throughout the cockroaches.

Approximately 2000 species of termites have been described and these are placed in six or seven families: Mastotermitidae, Hodotermitidae, Termopsidae (sometimes included in the Hodotermitidae), Kalotermitidae, Rhinotermitidae, Serritermitidae and the Termitidae. The latter family represents the higher termites and includes over 80% of all known termite species. *See* EXOPTERYGOTA; ORTHOPTERA.

The termite group is typically a tropical one, but certain genera do occur outside the tropics and may be found as far north as British Columbia and Ontario. The group is distinguished by the fact that all species are eusocial and all feed on cellulose. The castes, apart from the imagos and primary reproductives, are drawn from the larval forms. In this respect they differ from the Hymenoptera where the castes are all variants of adults. Termites live in nests of varying degrees of complexity, ranging from large exigeous mounds to diffuse or temporary galleries in wood or soil. *See* HYMENOPTERA.

Anatomical characteristics. The mature termite (alate or imago) has membranous wings which extend beyond the end of the abdomen. There is a pair of compound eyes, and a pair of ocelli is present in most groups. The wings are superimposed flat on the abdomen when the insect is not in flight. Flight is weak and fluttering and is usually short. When the alate alights, the wings are shed along a basal suture with the base of each wing (the wing scale) being retained. The alates vary in color from yellow, through brown, to coal black. Some species (usually lightly pigmented) fly during the night; others (usually heavily pigmented) fly during the day. The time of flight varies from species to species both with respect to the season of the year and the time of the day or night.

Soldier caste. In almost all termite species a second type of individual is produced in the colony. This is the soldier, which lacks wings, is nonreproductive, and is variously modified for defense. There are four rather distinct types of soldiers: mandibulate, phragmotic, nasutoid, and nasute. In mandibulate soldiers the head and mandibles are greatly enlarged and heavily sclerotized (**Fig. 1***a*). The mandibles may be biting, snapping, or pincherlike and more or less symmetrical or strongly asymmetrical. In phragmotic soldiers the mandibles are not as conspicuously enlarged as in mandibulate forms. The head is high and truncate in front and is used to plug openings in the workings.

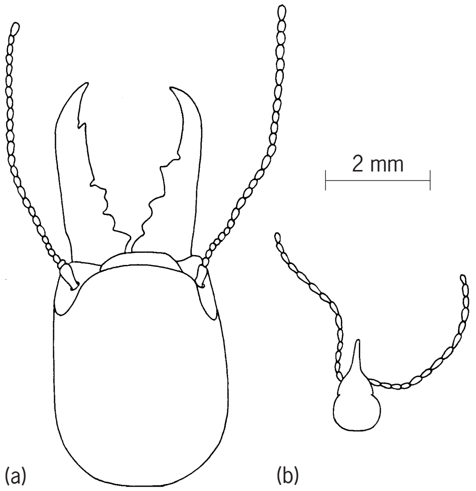


Fig. 1. Head capsule and antennae of (*a*) mandibulate soldier of *Zootermopsis nevadensis* showing enlarged mandibles, and (*b*) nasute soldier of *Lacessititermes palawanensis*.

In the families Rhinotermitidae and Termitidae there is a cephalic gland which opens via a small pore on the dorsal head surface. In some groups this gland and its opening have been variously modified for defense in the soldiers. In some rhinotermitid soldiers the opening of the gland lies at the anterior margin of the head, and the fluid is discharged into an open troughlike structure which extends forward from the head capsule. These have been termed nasutoid soldiers. Certain species have both mandibulate and nasutoid forms. Finally, in the termitid subfamily Nasutitermitinae the cephalic gland opens at the tip of an elongated tube or nasus which projects anteriorly, giving the head the appearance of a pear-shaped syringe (Fig. 1*b*). These are the nasute soldiers.

Soldiers are preceded during development by an intermediate stage (the soldier-nymph or white soldier) which is soldierlike in form but is unsclerotized. In general, mandibulate soldiers are rather large and occur in relatively small numbers in the colonies, whereas nasute soldiers are relatively minute and constitute as much as 30% of the population.

Worker caste. In the more advanced termites there is a third caste, the worker. True workers usually have some pigmentation as opposed to the immature termites, which are generally white. Workers lack wings, are nonreproductive, and have mandibles which resemble those of the imagos; they are usually blind. In many lower termites there is no distinct worker caste, and the work of the colony is performed by young individuals which have undergone at least two stages of development. These are still capable of becoming either alates or soldiers, but may continue to molt without becoming differentiated. Eventually they may undergo stationary molts with no appreciable change in size or structure. These “stabilized” individuals, which function as workers but are still capable of differentiation into other castes, have been termed pseudergates or pseudoworkers.

Replacement reproductives. In addition to the definitive castes (alate, soldier, and worker) another type of individual may occur in the colony under certain circumstances. These individuals are the supplementary or replacement reproductives. Although the original pair (king and queen) may live for two or three decades, the life of the colony itself is not limited by their survival. If one or both are lost, other individuals in the colony become reproductive. If these new reproductives have wing buds, they are termed nymphoid (or brachypterous or second-form) reproductives. If they lack wing buds, they are termed ergatoid (or apterous or third-form) reproductives.

Physiology. Two main areas of physiology which are of particular interest in Isoptera are nutrition and the development of the different castes.

Nutrition. Cellulose is the prime source of nutrition in termites, and may be obtained in nature as wood, humus in soil, or as grass. This consumption of cellulose, while most beneficial in natural areas, especially in the tropics, has caused some of the wood and grass eating forms to be considered pests.

The cellulose is masticated and passed through the gut to the hind intestinal caecum where it is digested by anaerobic protozoa (in the lower termites) and by bacteria (in the higher termites). Most researchers now believe that the cellulose is degraded to glucose and then to acetate which is able to pass through the chitinous lining of the hindgut where the large glucose molecule could not. The older idea that the glucose in the hindgut passes forward through a valve to the midgut for absorption by active transport is not now supported.

Nitrogen is obtained and conserved by both cannibalism and by adventitious consumption of fungi. Recently it has also been demonstrated that some of the symbiotic bacteria are able to fix atmospheric nitrogen. Uric acid is conserved and plays an important role as a nitrogen source.

Caste development. In many families of termites when the functional reproductives are removed from a colony some individuals that remain can transform into egg-laying and sperm-producing forms while retaining larval character; these are the replacement (or neotenic) reproductives. In 1932 it was proposed that the functional reproductives produce a chemical substance (pheromone) that inhibits the development of the gonads of the larvae. The hypothesis was elaborated over the years to include the involvement of twelve pheromones in the caste regulation of *Kalotermes* and *Zootermopsis*. This complex hypothesis, however, was quite speculative, as none of the postulated pheromones has yet been identified nor their origin discovered. Work by other investigators has discounted many of the earlier experiments. It is known, however, that vitellogenesis is affected by the presence of reproductives and that juvenile hormone plays an important part in the regulation; the mandibular glands may be involved. There probably is an inhibitory pheromone that requires contact or ingestion. *See* NEOTENY; PHEROMONE.

Soldier development is more complex. All that is known about it is that injection of a high titer of juvenile hormone into the larvae of some species will cause soldiers to form.

Biology. Eggs laid by a female reproductive hatch into mobile six-legged larvae which are broadly similar in form to the adult. They undergo successive molts over time and the later stages have external wing buds. The wing-budded forms finally molt into winged imagos with dark sclerotized bodies and well developed eyes. These imagos (alates) are produced at certain times of the year, and under certain weather conditions leave the nest all at once in a dispersal flight. The alates fly poorly and eventually settle on the ground or vegetation. Most species then lose their wings by actively twisting their bodies and breaking each wing along a basal suture line. The imagos with short wing stubs, now known as dealates, form into tandems, that is, a male following a female, and move to crevices in soil or wood. There they begin to excavate a small cavity, the nuptial cell or copularium, where mating takes place. The male and female are now referred to as the king and queen, and unlike the hymenopterans the male persists in the colony. Eggs are laid, hatch, and the new larval individuals are the nucleus of a new colony.

Not all larvae in a colony, however, become alate imagos. In most termites larvae may transform into soldiers and workers, while in many species larval forms may become sexually mature without becoming true adults. The main castes found in the Isoptera are these larval forms and, in this order particularly, they can be thought of as end points in individual development: once a termite transforms into a certain caste it loses ability to molt again. The soldier is designed for active defense while the worker is for excavation, feeding, and caring for other castes and individuals. As described above, some genera lack soldiers, some have more than one type of soldier, and some lower termites lack true workers.

While new colonies may be formed by imagos after a flight, in many genera new colonies can also be formed by a process known as sociotomy. This occurs when part of a colony is cut off from the main group by chance, or by the colony having grown so big as to stress the communication system that holds it together, or, as happens in some genera, by the active emigration of part of a colony.

All termites consume cellulose, but in some instances the source is not dead wood. Many termites, such as *Cornitermes*, ingest humus; the harvesting termites (for example, *Hodotermes* in Africa and *Drepanotermes* in Australia) collect and consume grass, while the spectacular *Macrotermes* of Africa construct fungus gardens.

Behavior. As in all social insects, communication is essential to their behavior. Colony cohesion and recognition, defense, construction of galleries and nests, pairing, and mating all are mediated through communication. Visual communication occurs only in adults, where the eyes are well developed; in other castes, where the eyes are rudimentary, communication is by chemical and tactile means. *See* ANIMAL COMMUNICATION; CHEMICAL ECOLOGY; SOCIAL INSECTS.

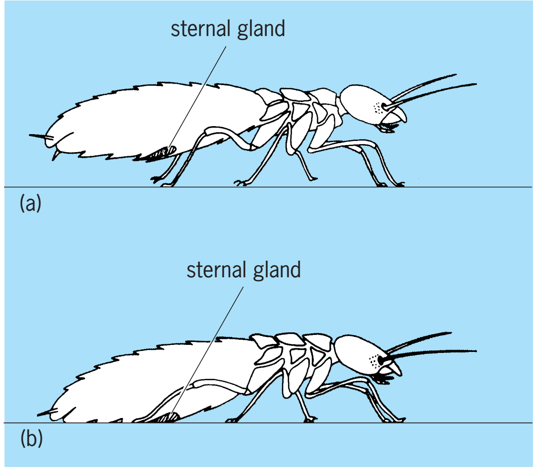


Fig. 2. Posture of a *Zootermopsis* nymph (*a*) during normal activity and (*b*) while laying a trail. (*After K. Krishna and F. Weesner, eds., Biology of Termites, vol. 1, 1969*)

Alarm and defense. Many termites are known to communicate alarm by a combination of characteristic tactile movements coupled with the laying of a chemical trail. The trail has its origin in the secretion of a pheromone by the sternal gland or glands of the abdomen (**Figs. 2** and **3**). A greater or lesser number of termites, depending on the intensity of the initial excitation, are recruited to the spot where the original disturbance occurred by following the trail after having been alerted by physical contact with an excited individual or individuals. This type of communication has also been found in other social insects and has been termed alarm recruitment.

Once a termite has been recruited to a specific locus by this mechanism, the action it then takes is dictated by factors at the site of disturbance. If the cause of alarm is a break in the nest or gallery, then building (in the form of fecal material deposition and the placement of debris) will occur by workers, while soldiers guard the site. If an intruder was the cause, then snapping will be exhibited by workers and mandibulate soldiers; soldiers of the Nasutitermitidae will eject a sticky terpene-containing substance from a gland situated in their heads. When the intruder is immobilized or killed, less excitation is generated and the termites' response is to bury it by depositing fecal material and debris as in a gallery break.

The response of a colony to alarm is further tuned by the phenomenon of polyethism, that is, the differential response of castes, and of individuals of differing ages within a caste, to the same stimulus. This phenomenon is seen most spectacularly in many species of *Nasutitermes* where in a localized disturbance there is an initial recruitment of soldiers as described above, but at the same time the workers retreat. In this case polyethism occurs partly in response to the frontal gland secretion of the soldiers, which has been shown to act as a short-lived recruitment pheromone for other soldiers in addition to its immobilizing function on intruders, and partly in the response to the sternal gland pheromone. The latter pheromone, when combined with the communication of other information such as the presence of food, is responsible for exciting the workers. *See* PHEROMONE.

Colony and kin recognition. Communication of kinship and membership of a colony is known to involve an odor characteristic of the colony. This odor is a composite one with contributions from pheromones, the recognition odors of various castes, the environment, and food. Intruders, including conspecifics from other colonies, are recognized and attacked when their odor is detected and does not match the generalized acquired colony odor. Colony odor is important in identifying aliens, in intraspecific competition, and in enhancing the degree of relationship between colony members by favoring inbreeding. The maintenance of a high degree of inbreeding would in turn favor the evolution of sociality by kin selection in these social insects with normal diploid sex determination.

Pairing and mating. Following the dispersal flight, pairing of male and female imagos takes place without copulation; this occurs after the nuptial cell has been formed. Typically a female de-alate will take up a calling position with the tip of her abdomen pointing vertically, exposing the sternal gland. (In imagos this gland has been shown to produce an excitatory pheromone for males.) A male bumping into a calling female becomes very excited and will tend to follow another moving termite. Upon being bumped, the female lowers her abdomen and begins running and the male follows her in tandem. If the male loses antennal contact with the female she will stop and resume the calling position while he moves randomly; once more contacting the female the process is repeated. Excavation occurs when the pair reaches wood or a crevice. In certain African termites pairing has a distinct visual component: before losing their wings, the females alight on stems of grass and flutter their wings in a typical display which attracts the males.

Mating takes place in the nuptial cell and occurs at intervals throughout the life of the reproductives. Copulation occurs in a rear-to-rear position.

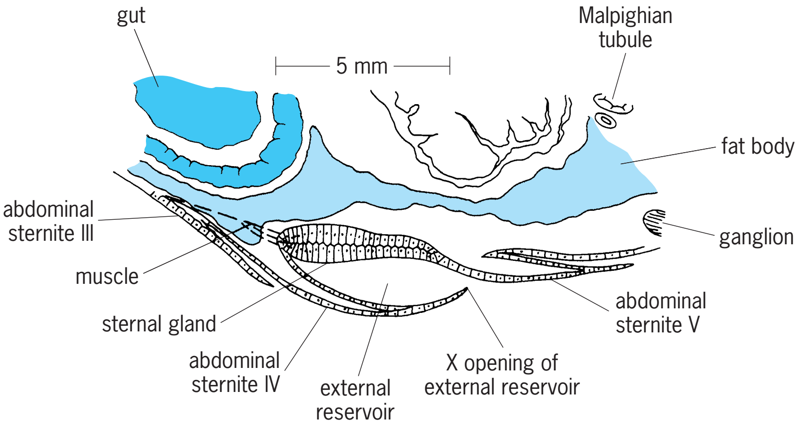


Fig. 3. Section through the abdomen of a nymph of *Zootermopsis nevadensis*, showing the sternal gland. (*After K. Krishna and F. Weesner, eds., Biology of Termites, vol. 1, 1969*)

Construction behavior. Building behavior has many manifestations, from the construction of simple galleries to the construction of the large bisymmetrical "magnetic" mounds made by *Amitermes meridionialis* in northern Australia. These mounds take up a north-south longitudinal axis which may occur through differential building activity in relation to a heat stimulus, as has been shown to occur with another termite species in the laboratory. The shape and orientation have evolved as a temperature regulating system in a termite that is unable to descend into the soil to avoid the heat (as other members of its genus do) because it is found in a habitat that is seasonally inundated.

For all termites, the basic act of building occurs as a response to a specific external stimulus. In some instances this primary building results in construction which, through its particular shape or chemical composition or both, acts as a new stimulus for further construction; this phenomenon is called "stigmergy." The building will stop when the results of building eventually reduce the original stimulus that initiated the building. This behavioral feedback applies in this situation and in alarm recruitment.

Evolution of sociality. Much has been proposed to explain the evolution of social behavior in the insects, and sociobiologists have heavily endorsed the idea that the haplodiploid form of sex determination found in the Hymenoptera is the basis of insect sociality. The termites, however, are diploid in both sexes and so destroy the generality of the hypothesis. Cytological studies, nonetheless, have inferred an exceptionally close genetic relationship between members of a termite colony, and this could partially explain the evolution of sociality and the concomitant altruistic behavior of nonsexual forms.

Probably several factors were involved in the evolution of sociality in termites, including the need to nurture the cellulose-degrading fauna of the hindgut. The similarity of the gut fauna of the wood-eating cockroach *Cryptocercus* to that of *Zootermopsis* supports the importance of diet specificity in the evolution of the social termites from a cockroach ancestor. *See* INSECTA; SOCIOBIOLOGY.

Alastair M. Stuart

Bibliography. K. Krishna and F. M. Weesner (eds.), *Biology of Termites*, vol. 1, 1969, vol. 2, 1970; A. M. Stuart, The role of pheromones in the initiation of foraging, recruitment and defence by soldiers of a tropical termite, *Nasutitermes corniger* (Motschulsky), *Chem. Senses*, 6:409–420, 1981.

Isopycnic

The line of intersection of an atmospheric isopycnic surface with some other surface, for instance, a surface of constant elevation or pressure. An isopycnic surface is a surface in which the density of the air is constant. Since specific volume is the reciprocal of density, isosteric surfaces coincide with isopycnic surfaces. On a surface of constant pressure, isopycnics coincide with isotherms, because on such a surface, density is a function solely of temperature. On a constant-pressure surface, isopycnics lie close together when the field is strongly baroclinic and are absent when the field is barotropic. *See* BAROCLINIC FIELD; BAROTROPIC FIELD; SOLENOID (METEOROLOGY).

Frederick Sanders; Howard B. Bluestein

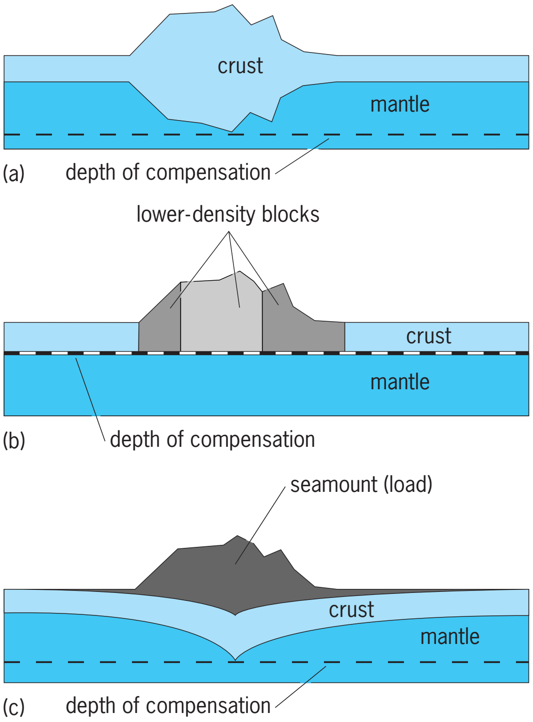
Bibliography. J. R. Holton, *An Introduction to Dynamic meteorology*, 4th ed., 2004; G. Visconti, *Fundamentals of Physics and Chemistry of the Atmosphere*, 2001.

Isostasy

The application of Archimedes' principle to the layered structure of the Earth. The elevated topography of Earth is roughly equivalent to an iceberg that floats in the surrounding, denser water. Just as an iceberg extends beneath the exposed ice, the concept of isostasy proposes that topography is supported, or compensated, by a deep root. The buoyant outer shell of the Earth, the crust, displaces the denser, viscous mantle in proportion to the surface elevation. Isostasy implies the existence of a level surface of constant pressure within the mantle, the depth of compensation. Above this surface the mass of any vertical column is equal. Equal pressure at depth can also be achieved by varying density structure or by the regional deflection of the lithosphere. *See* ARCHIMEDES' PRINCIPLE; EARTH CRUST.

This theory was independently proposed to explain inconsistencies in geodetic observations made by P. Bouguer (1735) in Peru and G. Everest (1836) in India. Each set of observations depended critically on the determination of vertical from a suspended plumb bob. In the absence of mass anomalies or topography, the bob will orient itself in the local direction of gravity, perpendicular to the surface of the Earth. Early geodesists recognized that mountain ranges such as the Andes and the Himalayas should deflect the plumb bob toward the mountains. Subsequent, independent work by Pratt (1855) demonstrated that the deflection was less than predicted by calculations, indicating that the gravitational attraction of the topography was somehow diminished. One explanation that was advanced (Airy isostasy; **illus.** *a*) for this discrepancy advocated a large iceberglike root of constant density, the mirror image of the surface topography. An alternative view (Pratt isostasy; illus. *b*) suggested that the lower interface was horizontal and that the density of the mountains was less than the density of the surrounding level regions. These theories, advanced in 1855, served to reduce errors in geodetic surveys, but they were unconfirmed until the direct measurement of crustal thickness beneath mountain ranges using seismic refraction techniques began in the late 1920s. These studies of crustal thickness have confirmed that mountains are being underlain by thick roots. *See* EARTH, GRAVITY FIELD OF; GEODESY.

Local isostasy achieves equilibrium directly beneath a load by varying either the density or thickness of that mass column. This model attributes no



Three major modes of isostatic compensation. (*a*) Airy isostasy, where the crustal density is constant beneath both the elevated topography and the level region; a large root extends beneath the elevated topography, and the depth of compensation is at the base of the crust where the pressure is constant. (*b*) Pratt isostasy, where the density of the crust varies inversely with the height of the topography; the depth of compensation is at the base of the horizontal crust–mantle boundary. (*c*) Flexural or regional isostasy, where the crust has some strength and is deflected beneath the elevated topography; the depth of compensation is a horizontal surface beneath the lowest extent of the crust.

inherent strength to the crust and assumes that the mantle is a simple fluid, redistributing mass to minimize pressure differences at depth. From studies of seamounts, oceanic trenches, foreland basins, and glacial rebound, it has become known that the outer shell of the Earth is rigid, responding to loads over a region broader than the load itself, and that the mantle is a viscous fluid with a time-dependent response to loads.

The simplest method of examining the response of the Earth is to study an area influenced by a discrete load such as a seamount or a continental glacier. If local isostasy (Pratt and Airy) is applicable, the region surrounding the load will be horizontal, unaffected by the load. In contrast, if the lithosphere has finite strength and regional or flexural isostasy (illus. *c*) is applicable, the surrounding regions will be deflected down toward the load. Gravity, bathymetry, and seismic studies of the crust surrounding Hawaii and other seamounts have demonstrated that the crust is downwarped beneath seamounts. The implication of this regional response is that the oceanic lithosphere has some strength and that the Earth's outer shell behaves elastically.

Just as seamounts can be considered loads on oceanic crust, so can continental ice sheets be viewed as loads on continental crust. Since the glacial retreat, Fennoscandia and Canada, which were directly beneath continental ice sheets, have experienced rapid uplift, as recorded by the series of beaches now resting tens to hundreds of meters above sea level. Uplift in this region directly beneath the previously loaded region confirms the prediction that the Earth's surface is in hydrostatic equilibrium, but does not discriminate between the local and regional isostasy. Beyond the furthest extent of the ice sheet, raised beaches are also found, indicating that the crustal deformation associated with the ice loading was regional rather than local. This regional deflection provides compelling evidence for lithospheric strength and regional isostasy. *See* EARTH; LITHOSPHERE. Robin E. Bell; Bernard J. Coakley

Bibliography. C. M. R. Fowler, *The Solid Earth: An Introduction to Global Geophysics*, 2d ed., 2005; R. Sabadini, K. Lambeck, and E. Boschi (eds.), *Glacial Isostasy, Sea-Level and Mantle Rheology*, 1991; B. J. Skinner and S. C. Porter, *The Dynamic Earth: An Introduction to Physical Geology*, 5th ed., 2003; A. B. Watts, *Isostasy & Flexure of the Lithosphere*, 2001; A. B. Watts et al., A multichannel seismic study of lithospheric flexure across the Hawaii-Emperor seamount chain, *Nature*, 313(6015):105–111, 1985.

Isotach

A line along which the speed of the wind is constant. Isotachs are customarily represented on surfaces of constant elevation or atmospheric pressure, or in vertical cross sections. The closeness of spacing of the isotachs is indicative of the intensity of the wind shear on such surfaces. In the region of a jet stream the isotachs are approximately parallel to the streamlines of wind direction and are closely spaced on either side of the core of maximum speed. *See* JET STREAM; WIND. Frederick Sanders; Howard B. Bluestein

Isothermal chart

A map showing the distribution of air temperature (or sometimes sea surface or soil temperature) over a portion of the Earth's surface, or at some level in the atmosphere. On it, isotherms are lines connecting places of equal temperature. The temperatures thus displayed may all refer to the same instant, may be averages for a day, month, season, or year, or may be the hottest or coldest temperatures reported during some interval.

Maps of mean monthly or mean annual temperature for continents, hemispheres, or the world sometimes show values reduced to sea level to eliminate the effect of elevation in decreasing average temperature by about 3.3°F/1000 ft (5.9°C/1000 m; see **illus.**). Such adjusted or sea-level maps represent the effects of latitude, continents, and oceans in modifying temperature; but they conceal the effect of mountains and highlands on temperature

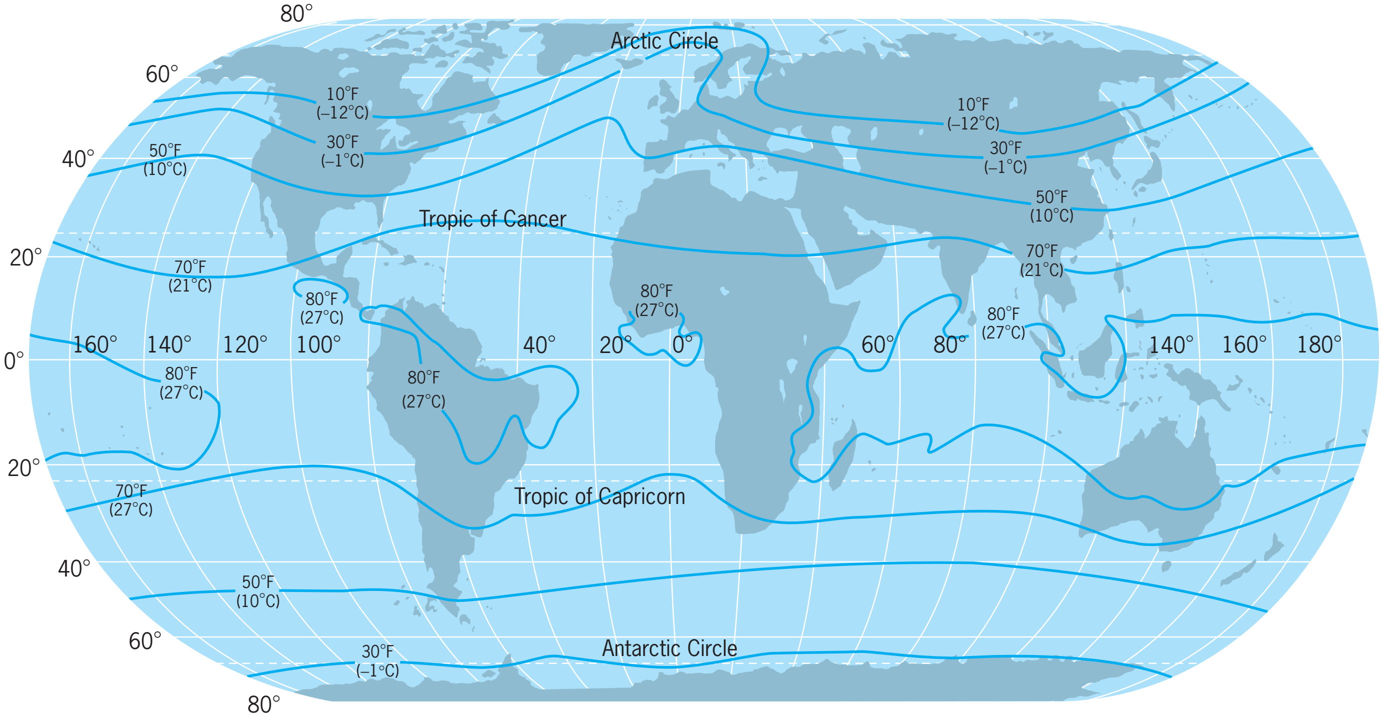


Chart showing selected January average isotherms. (*After D. K. Fellows, Our Environment*: *An Introduction to Physical Geography*, *2d ed.*, *John Wiley and Sons*, *1980*)

distributions. The first isothermal chart, prepared by Alexander von Humboldt in 1817 for low and middle latitudes of the Northern Hemisphere, was the first use of isopleth methods to show the geographic distribution of a quantity other than elevation.

These maps are now varied in type and use. Isothermal charts are drawn daily in major weather forecasting centers; 5-day, 2-week, and monthly charts are used regularly in long-range forecasting; mean monthly and mean annual charts are compiled and published by most national weather services, and are presented in standard books on, for example, climate, geography, and agriculture. *See* AIR TEMPERATURE; TEMPERATURE INVERSION; WEATHER FORECASTING AND PREDICTION.

Arnold Court

Isothermal process

A thermodynamic process that occurs with a heat addition or removal rate just adequate to maintain constant temperature. The change in internal energy per mole, U, accompanying a change in volume in an isothermal process is given by Eq. (1), where T

$$U_2 - U_1 = \int_{V_1}^{V_2} \left[T\left(\frac{\partial P}{\partial T}\right)_V - P \right] dV \tag{1}$$

is the temperature, P is pressure, and V is the volume per mole. The integral in Eq. (1) is zero for an ideal gas (which has the equation of state $PV = nRT$, where R is the gas constant and n the amount, so that the integrand vanishes), and for an incompressible condensed phase (solid or a liquid) for which the volume does not change with pressure. Thus, in both these cases, $U_2 = U_1$. For real gases and compressible liquids, the integral is nonzero, and the internal change is computed using a volumetric equation of state, that is, a relation between pressure, temperature, and volume.

The change in entropy, S, in an isothermal process is given by Eq. (2). For condensed phases with

$$S_2 - S_1 = \int_{V_1}^{V_2} \left(\frac{\partial P}{\partial T}\right)_V dV = -\int_{P_1}^{P_2} \left(\frac{\partial V}{\partial T}\right)_P dP \tag{2}$$

a small coefficient of thermal expansion, $S_2 \approx S_1$, while for real fluids the equation of state is used in Eq. (2) to obtain the numerical value of the entropy change. Equation (3) is used for the isothermal entropy change of an ideal gas.

$$S_2 - S_1 = nR \ln \frac{V_2}{V_1} = nR \ln \frac{P_2}{P_1} \tag{3}$$

In the case of a reversible isothermal process, the work done, W, by the fluid and the transferred heat to the fluid, Q, are given by Eqs. (4) and (5). The

$$W = -\int_{V_1}^{V_2} P dV = -Q + (U_2 - U_1) \tag{4}$$

$$Q = T(S_2 - S_1) \tag{5}$$

sign convention used here is that heat supplied to the fluid and work done on the fluid are positive. For a reversible isothermal process in an ideal gas, these equations reduce to Eq. (6). The comparable expres-

$$W = -Q = -nRT \ln \frac{V_2}{V_1} = +nRT \ln \frac{P_2}{P_1} \tag{6}$$

sions for real fluids, which depend on the equation of state, can be found in thermodynamics textbooks.

Phase transitions (that is, first-order phase transitions such as solid to liquid, or liquid to vapor) in a pure material occur at constant temperature and constant pressure according to the Gibbs phase rule. The Gibbs free energy change for a first-order phase transition, ΔG, is identically zero at the transition temperature, the enthalpy and entropy changes, ΔH and ΔS respectively, are related as given in Eq. (7).

$$\begin{aligned} &\Delta G = 0 = \Delta H - T\Delta S \\ \text{or}\quad &\Delta S = \frac{\Delta H}{T} = \frac{\Delta U + P\Delta V}{T} \end{aligned} \tag{7}$$

See CHEMICAL THERMODYNAMICS; GAS; PHASE RULE; PHASE TRANSITIONS; THERMODYNAMIC PROCESSES.

Stanley I. Sandler

Bibliography. S. I. Sandler, *Chemical and Engineering Thermodynamics*, 3d ed., 1999; R. E. Sonntag, C. Borgnakke, and G. J. Van Wylen, *Fundamentals of Thermodynamics*, 5th ed., 1997.

Isotone

One of two or more atoms which display a constant difference $A - Z$ between their mass number A and their atomic number Z. Thus, despite differences in the total number of nuclear constituents, the numbers of neutrons in the nuclei of isotones are the same. The numbers of naturally occurring isotones provide useful evidence concerning the stability of particular neutron configurations. For example, the relatively large number (six and seven, respectively) of naturally occurring 50- and 82-neutron isotones suggests that these nuclear configurations are especially stable. On the other hand, from the fact that most atoms with odd numbers of neutrons are anisotonic, one may conclude that odd-neutron configurations are relatively unstable. *See* NUCLEAR STRUCTURE.

Henry E. Duckworth

Isotope

One member of a (chemical-element) family of atomic species which has two or more nuclides with the same number of protons (Z) but a different number of neutrons (N). Because the atomic mass is determined by the sum of the number of protons and neutrons contained in the nucleus, isotopes differ in mass. Since they contain the same number of protons (and hence electrons), isotopes have the same chemical properties. However, the nuclear and atomic properties of isotopes can be different. The electronic energy levels of an atom depend upon the nuclear mass. Thus, corresponding atomic levels of isotopes are slightly shifted relative to each other. A nucleus can have a magnetic moment which can interact with the magnetic field generated by the electrons and lead to a splitting of the electronic levels. The number of resulting states of nearly the same energy depends upon the spin of the nucleus and the characteristics of the specific electronic level. *See* ATOMIC STRUCTURE AND SPECTRA; HYPERFINE STRUCTURE; ISOTOPE SHIFT.

Of the 108 elements reported thus far, 81 have at least one stable isotope whereas the others exist only in the form of radioactive nuclides. Some radioactive nuclides (for example, ^{115}In, ^{232}Th, ^{235}U, ^{238}U) have survived from the time of formation of the elements. Several thousand radioactive nuclides produced through natural or artificial means have been identified. *See* RADIOISOTOPE.

Of the 83 elements which occur naturally in significant quantities on Earth, 20 are found as a single isotope (mononuclidic), and the others as admixtures containing from 2 to 10 isotopes. Isotopic composition is mainly determined by mass spectroscopy. *See* MASS SPECTROSCOPE.

Nuclides with identical mass number (that is, $A = N + Z$) but differing in the number of protons in the nucleus are called isobars. Nuclides having different mass number but the same number of neutrons are called isotones. *See* ISOBAR (NUCLEAR PHYSICS); ISOTONE.

Nuclear stability. The stability of a nuclide is governed by its total energy E as given by the Einstein relation $E = Mc^2$, where M is the nuclidic mass and c is the velocity of light. If E is less than the combined energies of possible decay products, the nuclide will be stable. A major factor in determining stability is the relative strength of the nuclear force which acts to attract nucleons and the coulombic force (repulsive) which arises from the electric charge on the protons. Nuclides with an even number of protons or neutrons are prevalent in the **table**. Of the 287 nuclides tabulated, 168 are even-even (that is, an even number of both neutrons and protons), 110 are odd, and only 9 are odd-odd. This demonstrates the increased attraction of the nuclear force between pairs of nucleons of the same type (the pairing effect). Nuclides for which the number of either protons or neutrons (or both) comprises so-called magic numbers (for example, 8, 20, 50, 82, etc.) have increased stability. *See* NUCLEAR STRUCTURE.

Isotopic abundance. The term isotopic abundance refers to the isotopic composition of an element found in its natural terrestrial state. The isotopic composition for most elements does not vary much from sample to sample. This is true even for samples of extraterrestrial origin such as meteorites and lunar materials brought back to Earth by United States crewed and Soviet uncrewed missions. However, there are a few exceptional cases for which variations of up to several percent have been observed. There are several phenomena that can account for such variations, the most likely being some type of nuclear process which changes the abundance of one isotope relative to the others. For some of the lighter elements, the processes of distillation or chemical exchange between different chemical compounds could be responsible for isotopic differences. *See* NUCLEAR REACTION; RADIOACTIVITY.

The lead isotopes ^{206}Pb, ^{207}Pb, and ^{208}Pb are stable and are end products of naturally decaying ^{238}U,

Natural isotopic compositions of the elements

Atomic no.	Element symbol	Mass no.	Isotopic abundance, %
1	H*	1	99.985
		2	0.015
2	He*	3	0.000138
		4	99.999862
3	Li*	6	7.5
		7	92.5
4	Be	9	100
5	B*	10	19.9
		11	80.1
6	C*	12	98.90
		13	1.10
7	N*	14	99.634
		15	0.366
8	O	16	99.762
		17	0.038
		18	0.200
9	F*	19	100
10	Ne*†	20	90.51
		21	0.27
		22	9.22
11	Na	23	100
12	Mg	24	78.99
		25	10.00
		26	11.01
13	Al	27	100
14	Si	28	92.23
		29	4.67
		30	3.10
15	P	31	100
16	S*	32	95.02
		33	0.75
		34	4.21
		36	0.02
17	Cl	35	75.77
		37	24.23
18	Ar†	36	0.337
		38	0.063
		40	99.600
19	K	39	93.2581
		40	0.0117
		41	6.7302
20	Ca	40	96.941
		42	0.647
		43	0.135
		44	2.086
		46	0.004
		48	0.187
21	Sc	45	100
22	Ti	46	8.0
		47	7.3
		48	73.8
		49	5.5
		50	5.4
23	V	50	0.250
		51	99.750
24	Cr	50	4.35
		52	83.79
		53	9.50
		54	2.36
25	Mn	55	100
26	Fe	54	5.8
		56	91.72
		57	2.2
		58	0.28
27	Co	59	100
28	Ni	58	68.27
		60	26.10
		61	1.13
		62	3.59
		64	0.91
29	Cu	63	69.17
		65	30.83
30	Zn	64	48.6
		66	27.9
		67	4.1
		68	18.8
		70	0.6
31	Ga	69	60.1
		71	39.9
32	Ge	70	20.5
		72	27.4
		73	7.8
		74	36.5
		76	7.8
33	As	75	100
34	Se	74	0.9
		76	9.0
		77	7.6
		78	23.5
		80	49.6
		82	9.4
35	Br	79	50.69
		81	49.31
36	Kr†	78	0.35
		80	2.25
		82	11.6
		83	11.5
		84	57.0
		86	17.3
37	Rb	85	72.165
		87	27.835
38	Sr†	84	0.56
		86	9.86
		87	7.00
		88	82.58
39	Y	89	100
40	Zr	90	51.45
		91	11.27
		92	17.17
		94	17.33
		96	2.78
41	Nb	93	100
42	Mo	92	14.84
		94	9.25
		95	15.92
		96	16.68
		97	9.55
		98	24.13
		100	9.63
44	Ru	96	5.52
		98	1.88
		99	12.7
		100	12.6
		101	17.0
		102	31.6
		104	18.7
45	Rh	103	100
46	Pd	102	1.02
		104	11.14
		105	22.33
		106	27.33
		108	24.46
		110	11.72
47	Ag	107	51.839
		109	48.161
48	Cd	106	1.25
		108	0.89
		110	12.49
		111	12.80
		112	24.13
		113	12.22
		114	28.73
		116	7.49
49	In	113	4.3
		115	95.7
50	Sn	112	1.0
		114	0.7
		115	0.4
		116	14.7
		117	7.7
		118	24.3
		119	8.6
		120	32.4
		122	4.6
		124	5.6

Natural isotopic compositions of the elements (*cont.*)

Atomic no.	Element symbol	Mass no.	Isotopic abundance, %
51	Sb	121	57.3
		123	42.7
52	Te	71	39.9
		120	0.096
		122	2.60
		123	0.908
		124	4.816
		125	7.14
		126	18.95
		128	31.69
		130	33.80
53	I	127	100
54	Xe†	124	0.10
		126	0.09
		128	1.91
		129	26.4
		130	4.1
		131	21.2
		132	26.9
		134	10.4
		136	8.9
55	Cs	133	100
56	Ba	130	0.106
		132	0.101
		134	2.417
		135	6.592
		136	7.854
		137	11.23
		138	71.70
57	La	138	0.09
		139	99.91
58	Ce	136	0.19
		138	0.25
		140	88.48
		142	11.08
59	Pr	141	100
60	Nd	142	27.13
		143	12.18
		144	23.80
		145	8.30
		146	17.19
		148	5.76
		150	5.64
62	Sm	144	3.1
		147	15.0
		148	11.3
		149	13.8
		150	7.4
		152	26.7
		154	22.7
63	Eu	151	47.8
		153	52.2
64	Gd	152	0.20
		154	2.18
		155	14.80
		156	20.47
		157	15.65
		158	24.84
		160	21.86
65	Tb	159	100
66	Dy	156	0.06
		158	0.10
		160	2.34
		161	18.9
		162	25.5
		163	24.9
		164	28.2
67	Ho	165	100
68	Er	162	0.14
		164	1.61
		166	33.6
		167	22.95
		168	26.8
	Er	170	14.9
69	Ta	169	100
70	Yb	168	0.13
		170	3.05
		171	14.3
		172	21.9
		173	16.12
		174	31.8
		176	12.7
71	Lu	175	97.40
		176	2.60
72	Hf	174	0.16
		176	5.2
		177	18.6
		178	27.1
		179	13.74
		180	35.2
73	Ta	180	0.012
		181	99.988
74	W	180	0.13
		182	26.3
		183	14.3
		184	30.67
		186	28.6
75	Re	185	37.40
		187	62.60
76	Os†	184	0.02
		186	1.58
		187	1.6
		188	13.3
		189	16.1
		190	26.4
		192	41.0
77	Ir	191	37.3
		193	62.7
78	Pt	190	0.01
		192	0.79
		194	32.9
		195	33.8
		196	25.3
		198	7.2
79	Au	197	100
80	Hg	196	0.15
		198	10.1
		199	17.0
		200	23.1
		201	13.2
		202	29.65
		204	6.8
81	Tl	203	29.524
		205	70.467
82	Pb†	204	1.4
		206	24.1
		207	22.1
		208	52.4
83	Bi	209	100
90	Th	232	100
92	U*	234	0.0055
		235	0.7200
		236	99.2745

*Isotopic composition may vary with sample depending upon geological or biological origin.

†Isotopic composition may vary with sample because some of the isotopes may be formed as a result of radioactive decay or nuclear reactions.

^{235}U, and ^{232}Th, respectively, whereas ^{204}Pb is not produced by any long-lived decay chain. Thus, the isotopic composition of lead samples will depend upon their prior contact with thorium and uranium. The potassium isotope ^{40}K has a half-life of 1.28×10^9 years and decays by beta-ray emission to ^{40}Ar and by electron capture to ^{40}Ca, which are both stable. This can cause the argon in potassium-bearing minerals to differ in isotopic abundance from that found in air. It is possible to determine the age of rocks by measuring the ratio of their ^{40}K/^{40}Ar content. This ratio technique can also be used for rock samples which bear other long-lived, naturally occurring isotopes such as ^{87}Rb (rubidium), thorium, and uranium. *See* LEAD ISOTOPES (GEOCHEMISTRY); ROCK AGE DETERMINATION.

An interesting example of anomalous isotopic compositions has been observed in the Oklo uranium deposit in Gabon (western Africa). Based upon extensive research, it has been concluded that this is the site of a natural chain reaction that took place about 1.8×10^9 years ago. Much of the uranium in this formation has been depleted of the fissionable isotope ^{235}U. The isotopic composition of some of the other elements found at or near this deposit has also been altered as a result of fission, neutron absorption, and radioactive decay.

Use of separated isotopes. The areas in which separated (or enriched) isotopes are utilized have become fairly extensive, and a partial list includes nuclear research, nuclear power generation, nuclear weapons, nuclear medicine, and agricultural research. Various methods are employed to prepare separated isotopes. Mass spectroscopy is used in the United States and Russia to prepare inventories of separated stable isotopes. Distillative, exchange, and electrolysis processes have been used to produce heavy water (enriched in deuterium, ^{2}H), which is used as a neutron moderator in some reactors. The uranium enrichment of ^{235}U, which is used as a fuel in nuclear reactors, has mainly been accomplished by using the process of gaseous diffusion in uranium hexafluoride gas in very large plants. This method has the disadvantage of requiring large power consumption. Techniques which can overcome this problem and are finding increasing favor include centrifugal separation and laser isotope separation. *See* ISOTOPE SEPARATION; NUCLEAR REACTOR.

For many applications there is a need for separated radioactive isotopes. These are usually obtained through chemical separations of the desired element following production by means of a suitable nuclear reaction. Separated radioactive isotopes are used for a number of diagnostic studies in nuclear medicine, including the technique of positron tomography. *See* NUCLEAR MEDICINE.

Studies of metabolism, drug utilization, and other reactions in living organisms can be done with stable isotopes such as ^{13}C, ^{15}N, ^{18}O, and ^{2}H. Molecular compounds are "spiked" with these isotopes, and the metabolized products are analyzed by using a mass spectrometer to measure the altered isotopic ratios. The use of separated isotopes as tracers for determining the content of a particular element in a sample by dilution techniques had broad applicability. *See* ISOTOPE DILUTION TECHNIQUES; RADIOISOTOPE (BIOLOGY).

Atomic mass. Atomic masses are given in terms of an internationally accepted standard which at present defines an atomic mass unit (amu) as exactly equal to one-twelfth the mass of a neutral atom of ^{12}C in its electronic and nuclear ground states. On this scale the mass of a ^{12}C atom is equal to 12.0 amu. Atomic masses can be determined quite accurately by means of mass spectroscopy, and also by the use of nuclear reaction data with the aid of the Einstein mass-energy relation. Atomic weight is the average mass per atom per amu of the natural isotopic composition of an element. Practically all atomic weights are now calculated by taking the sum of the products of the fractional isotopic abundances times their respective masses in amu. The atomic weight of a neutral atom of a nuclide is nearly equal to an integer value A, because the mass of both the neutron and proton is almost identically 1 amu. *See* ATOMIC MASS.

Daniel J. Horen

Bibliography. A. P. Dickin, *Radiogenic Isotope Geology*, 1994, reprint 1997; H. E. Duckworth, *Mass Spectroscopy*, 2d ed., 1990; K. Lajtha and R. Michener, *Stable Isotopes in Ecology*, 1994; D. S. Schimel, *Theory and Application of Tracers*, 1993; H. P. Taylor, J. R. O'Neill, and I. R. Kaplan, *Stable Isotope Geochemistry*, 1991; R. R. Wolfe, *Radioactive and Stable Isotope Tracers in Biomedicine*, 1992.

Isotope dilution techniques

Analytical techniques that involve the addition to a sample of an isotopically labeled compound. Soon after the discovery of the stable heavy isotopes of hydrogen, carbon, nitrogen, and oxygen, their value in analytical chemistry was recognized. Stable isotopes were particularly useful in the analysis of complex mixtures of organic compounds where the isolation of the desired compound with satisfactory purity was difficult and effected only with low or uncertain yields. The addition of a known concentration of an isotopically labeled compound to a sample immediately produces isotope dilution if the particular compound is present in the sample. After thorough mixing of the isotopically labeled compound with the sample, any technique that determines the extent of the isotopic dilution suffices to establish the original concentration of the compound in the mixture. Isotope dilution techniques exploit the difficulty in the separation of isotopes, with the isotopically labeled "spike" following the analytically targeted compound through a variety of separation procedures prior to isotopic analysis. *See* ISOTOPE.

The technique depends on the availability of a stable or radioisotope diluent with isotope abundance ratios differing markedly from those of the naturally occurring elements. With monoisotopic elements, such as sodium or cesium, radioactive elements of sufficiently long life can be used in isotope dilution techniques. In 1954 M. G. Ingraham identified 68 elements that were analyzable by isotope dilution techniques. The list has since been expanded considerably.

Biochemical applications. The original applications of isotope dilution were by biochemists interested in complex mixtures of organic compounds. In these studies care had to be taken to ensure the stability of the labeled compound and its resistance to isotopic exchange reactions. Nitrogen-15–labeled glycine for example, could be used to determine glycine in a mixture of amino acids obtained from a protein. Deuterium-labeled glycine could not be used reliably if the deuterium isotopes were attached to the glycine amino or carboxyl group, because in these locations deuterium is known to undergo exchange reactions with hydrogens in the solvent or in other amino acids. Deuterium is very useful in elemental isotopic analysis where total hydrogen or exchangeable hydrogen concentrations are desired. *See* BIOCHEMISTRY; DEUTERIUM.

High-sensitivity applications. Applications of isotope dilution techniques have also been found in geology, nuclear science, and materials science. These applications generally focus on the very high sensitivity attainable with these techniques. Isotopes of argon, uranium, lead, thorium, strontium, and rubidium have been used in geologic age determinations of minerals and meteorites. Taking the estimated error as a measurement of sensitivity, isotopic dilution analyses of uranium have been done down to 4 parts in 10^{12} and on thorium to 8 parts in 10^{9}. Studies in geology and nuclear science require the determination of trace amounts of radiogenic products. If the half-life and decay scheme of the parent nuclide is known, then isotopic dilution determinations of parent and daughter isotopes provide a basis for the calculation of the age of the sample. If the age or history of the sample is known, then determination of the trace concentrations of isotopes provides information on pathways of nuclear reactions. *See* RADIOACTIVITY; ROCK AGE DETERMINATION.

An example of the latter type of investigation is the determination of yields of cesium (Cs) from bombardment of gold or uranium with very high-energy protons. The use of radioactive ^{131}Cs with techniques of high-sensitivity mass spectrometry permitted the determination of stable and long-lived Cs isotopes in concentrations of less than 10^{-15} g. The fundamental limitation in the sensitivity of isotope dilution as a trace analytical technique is in contamination of reagents used in the analytical procedure. In the case of the study on cesium cited above, reagent water prepared by triple distillation in quartz columns was found to have approximately 10^{-15} g of natural ^{133}Cs per cubic centimeter.

Studies using isotope dilution techniques have dealt with trace analysis of toxic materials in biological systems. Uranium-233 has been used as an isotopic spike in the determination of uranium in bovine liver and in oyster tissue. Assay of 10^{-10} g of uranium with a precision of better than 0.5% was achieved. Contamination of reagents remains the limiting factor in sensitivity, with water assayed at the 4.5–22-femtogram level by isotope dilution. Uranium is an example of an element that is a potent biotoxin as well as a radiological hazard. Its assay in trace concentrations is of considerable importance in toxicology. *See* RADIATION BIOLOGY; TOXICOLOGY.

Lewis Friedman

Bibliography. J. H. Chen and G. J. Wasserburg, Isotopic determination of uranium in picomole and subpicomole quantities, *Anal. Chem.*, 53:2060–2067, 1981; M. G. Ingraham, Stable isotope dilution as an analytical tool, *Annu. Rev. Nucl. Sci.*, 4:81–92, 1954; J. A. Jonckheere, A. P. DeLeenheer, and H. L. Steyaert, Statistical evaluation of calibration curve nonlinearity in isotope dilution gas chromatography/mass spectrometry, *Anal. Chem.*, 55:153–155, 1983; J. F. Pickup and K. McPherson, Theoretical considerations in stable isotope dilution mass spectrometry for organic analysis, *Anal. Chem.*, 48:1885–1890, 1976; D. C. Reamer and C. Veillon, Determination of selenium in biological materials by stable isotope dilution gas chromatography-mass spectrometry, *Anal. Chem.*, 53:2166–2169, 1981.

Isotope separation

The physical separation of different isotopes of an element from one another. The different isotopes of an element as it occurs in nature may have similar chemical properties but completely different nuclear reaction properties. Therefore, nuclear physics and nuclear energy applications often require that the different isotopes be separated. However, similar physical and chemical properties make isotope separation by conventional techniques unusually difficult. Fortunately, the slight mass difference of isotopes of the same element makes separation possible by using especially developed processes, some of which involve chemical industry distillation concepts. *See* ISOTOPE.

Isotope separation depends on the element involved and its industrial application. Uranium isotope separation has by far the greatest industrial importance, because uranium is used as a fuel for nuclear power reactors, particularly the predominant light-water reactor type. The two main isotopes found in nature are ^{235}U and ^{238}U, which are present in weight percentages (w/o) of 0.711 and 99.283, respectively. Trace amounts of ^{234}U are also present in natural uranium. In order to be useful as a fuel for light-water reactors, the weight percentage of ^{235}U must be increased to between 2 and 5. The process of increasing the ^{235}U content is known as uranium enrichment, and the process of enriching is referred to as performing separative work. *See* NUCLEAR FUELS; NUCLEAR REACTOR; URANIUM.

The production of heavy water is another example of isotope separation. Because of its exceptional neutron slowing-down characteristics, heavy water is used as a moderator in some types of natural and very low enriched uranium-fueled nuclear reactors. Heavy water is obtained by isotope separation of light hydrogen (^{1}H) and heavy hydrogen (^{2}H) in natural water. Heavy hydrogen is usually referred to as deuterium (D). All natural waters contain ^{1}H and ^{2}H, in concentrations of 99.985 and 0.015 w/o, respectively, in the form of H_2O and D_2O (deuterium oxide). Isotope separation increases the concentration of the D_2O, and thus the purity of the heavy water. *See* DEUTERIUM; HEAVY WATER.

The development of laser isotope separation technology provided a range of potential applications. For example, enrichment of natural gadolinium in the isotope ^{157}Gd could result in the production of very high quality material for reactivity control of light-water reactors. Other isotope separation applications range from space-flight power sources (^{238}Pu) to medical magnetic resonance imaging (^{13}C) and medical research (^{15}O).

Separation technologies. The isotope separation process that is best suited to a particular application depends on the state of technology development as well as on the mass of the subject element and the quantities of material involved. Processes such as electromagnetic separation which are suited to research quantities of material are generally not suited to industrial separation quantities. However, the industrial processes that are used, gaseous diffusion and gas centrifugation, are not suited to separating small quantities of material.

Uranium isotope separation by either the gaseous diffusion or gas centrifuge process requires that the uranium be in the form of uranium hexafluoride (UF_6) gas. Both separation processes take advantage of the slight difference in the mass of the ^{235}U and ^{238}U atoms. For gaseous diffusion, the separator stage may consist of a single gas diffuser, gas compressor, gas cooler, compressor drive motor, and associated piping and valving, while a gas-centrifuge stage may consist of many centrifuge machines and associated hardware operating in parallel. In either case, the unit or units operating in parallel on material of the same mole fraction form a stage. Each stage has feed, enriched product, and depleted waste streams and a characteristic stage-separation efficiency factor. Stages connected in series to obtain progressive separation are referred to as cascades.

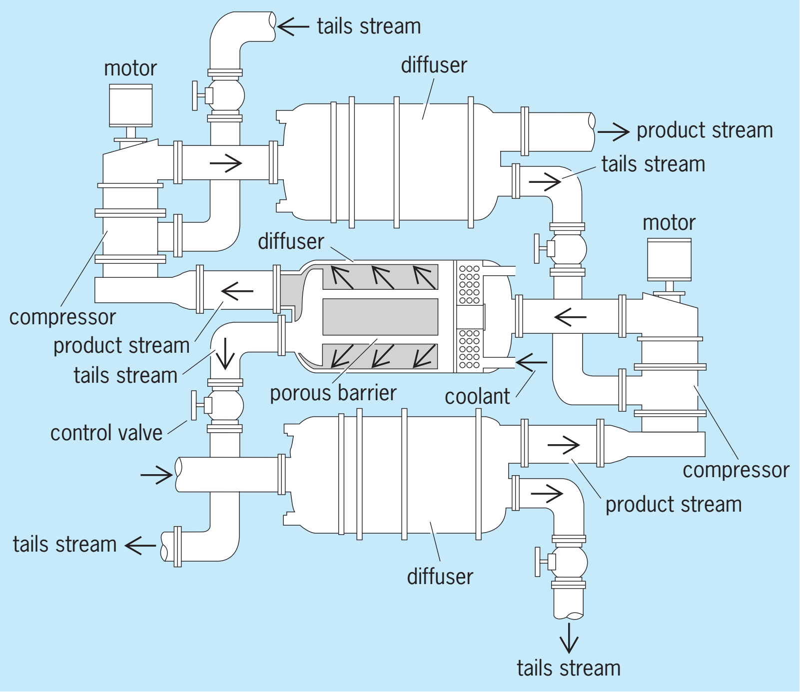


Fig. 1. Gaseous diffusion stage. (*After U.S. Atomic Energy Commission Report ORO-684, January 1972*)

In the course of accomplishing the separation of ^{235}U and ^{238}U in each of the separation stages (either diffusion or centrifuge), large amounts of uranium hexafluoride gas are recirculated. Application of chemical engineering theory using abundance ratios, reflux ratios, and flow-rate considerations, analogous to distillation theory, yields a term for the total flow of uranium in the separation cascade consisting of the product of two terms. One term is a function of the stage-separation factor, which is dependent on the molecular weights of the isotopes being separated, and the other is a function of the flow rates and isotopic composition of the feed, product, and waste. This second term is called separative work (SW), and is usually referred to in the following units: either kgU SWU or SWU in the United States, and either kgSW or tonnes SW in other countries.

Gaseous diffusion. In the gaseous diffusion process, uranium hexafluoride gas is circulated under pressure through specially designed equipment containing porous barriers. The barriers contain hundreds of millions of pores per square inch, with the average diameter of the pores being only a few millionths of an inch. The kinetic energies of the uranium hexafluoride molecules depend only on the temperature, which is the same within each stage. Since the kinetic energy of a molecule is proportional to the product of its mass and the square of its speed, the lighter molecules (^{235}U) have higher speeds causing them to strike the barrier more frequently than the heavy molecules (^{238}U), and therefore they have a higher probability of penetration. Thus, the lighter molecules (^{235}U) diffuse (or more correctly, effuse) through the barriers faster than the heavy molecules (^{238}U), with the result that the downstream side of the barrier is enriched in the lighter gas and the upstream side is depleted in the lighter gas. *See* DIFFUSION; KINETIC THEORY OF MATTER.

The stage-separation factor for the gaseous diffusion process depends only on the molecular weights of $^{235}U^{19}F_6$ and $^{238}U^{19}F_6$, which are 349 and 352, respectively; it is equal to 1.0043. Because the maximum possible value of the separation factor is so close to unity, a gaseous diffusion plant typically needs a very large number of diffuser-compressor-cooler units (**Fig. 1**) connected in series in order to obtain sufficiently enriched product. For example, approximately 1400 stages are needed to produce the 4 w/o ^{235}U that is typical of that used in light-water reactors. The throughput in each stage of a large diffusion plant can require large compressors driven by motors of several thousand horsepower (several megawatts).

Gas centrifugation. The gas centrifuge process has the advantage that its isotope separation capability for uranium is proportional to the absolute mass difference of the uranium hexafluoride molecules, rather than the relative difference, which is the case for gaseous diffusion. The separation effect is achieved by pressure diffusion through the gravitational field generated within the centrifuge rotor. The concentration of the heavy molecule tends to increase toward the rotor wall, while that of the lighter molecule tends to increase at the center of the rotor.

The radial separation factor of a gas centrifuge machine depends on the angular velocity of rotation, the radius of the rotor wall, and the absolute gas temperature. For a rotor radius of 6 cm (2.4 in.), a rotational speed of 40,000 revolutions per minute (corresponding to a peripheral speed of 250 m/s or 800 ft/s), and a gas temperature of 300 K (80°F), the separation factor is 1.0387; the difference of this separation factor from unity is nine times that of the separation factor for the gaseous diffusion process. The separation capability is approximately inversely proportional to the absolute temperature, and operating at a low or at least moderate temperature is therefore desirable. However, an adequate temperature margin must always be maintained to ensure that the uranium hexafluoride process gas does not condense.

In a practical centrifuge, an internal countercurrent circulation flow is introduced (**Fig. 2**). This flow has the effect of generating an axial concentration difference which is a factor larger than the elementary radial separation factor discussed above. Thus, even a centrifuge of very modest speed has a separation factor greatly exceeding that of the gas diffusion process. The countercurrent flow also permits the enriched and depleted fractions to be removed at the end of the rotor. In practice, centrifuge throughput has a strong dependence on the rotor peripheral speed and length. However, rotor material composition limits peripheral speed since the rotor will break if induced stresses exceed the tensile strength of the material. *See* CENTRIFUGATION.

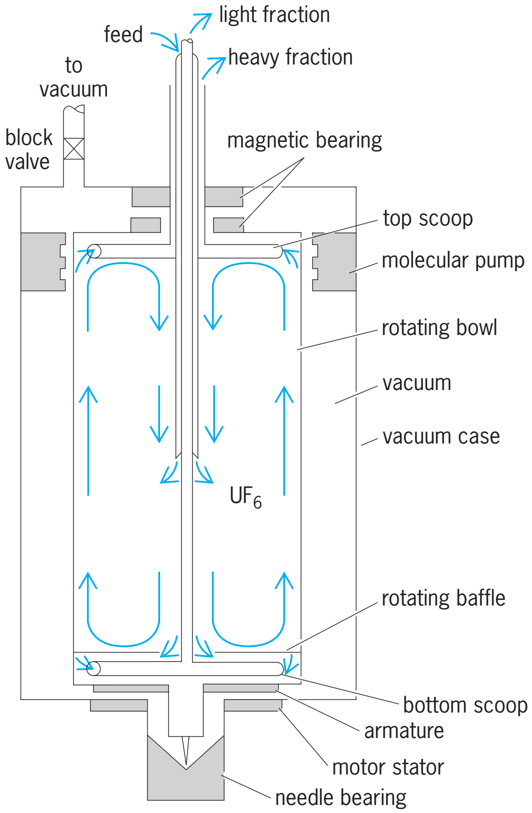


Fig. 2. Gas centrifuge machine. (*After M. Benedict, T. H. Pigford, and H. W. Levi, Nuclear Chemical Engineering, 2d ed., McGraw-Hill, 1981*)

Laser separation. Three experimental laser isotope separation technologies for uranium are the atomic vapor laser isotope separation (AVLIS) process, the uranium hexafluoride molecular laser isotope separation (MLIS) process, and the separation of isotopes by laser excitation (SILEX) process, which was announced in Australia in the early 1990s. The AVLIS process, which is more experimentally advanced than the MLIS and SILEX processes, exploits the fact that the different electron energies of ^{235}U and ^{238}U absorb different colors of light (that is, different wavelengths). In the AVLIS process, lasers are tuned to emit a combination of colors that will be absorbed only by a ^{235}U atom, which then emits an electron. This photoionization leaves the ^{235}U atom with a net positive charge, allowing the atom, now an ion, to be selectively separated by using electromagnetic fields. AVLIS technology is inherently more efficient than either the gaseous diffusion or gas centrifuge processes. It can enrich natural uranium to 4 w/o ^{235}U in a single step. In the United States, the AVLIS process is being developed to eventually replace the gaseous diffusion process for commercially enriching uranium. *See* PHOTOIONIZATION.

The AVLIS process includes two major component systems: a laser system and a separator system (**Fig. 3**). The laser system uses two types of lasers: dye lasers that generate the light used for photoionization of the uranium, and copper-vapor lasers that energize the dye lasers. Powerful green-yellow light from electrically driven copper-vapor lasers is converted to red-orange light in the dye laser. This red-orange light is tuned to the precise colors that are absorbed by ^{235}U but not by ^{238}U. This laser–pumped-laser scheme is required because dye lasers cannot be powered directly by electricity. *See* LASER.

In the separator system, uranium metal is melted and vaporized by means of an electron beam that

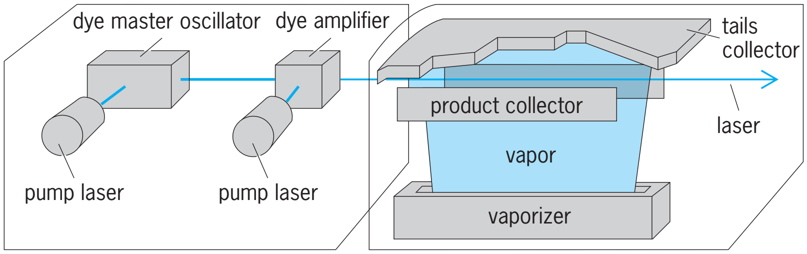


Fig. 3. Atomic vapor laser isotope separation (AVLIS) process. (*After Lawrence Livermore National Laboratory Report LLL-TB-072, 1985*)

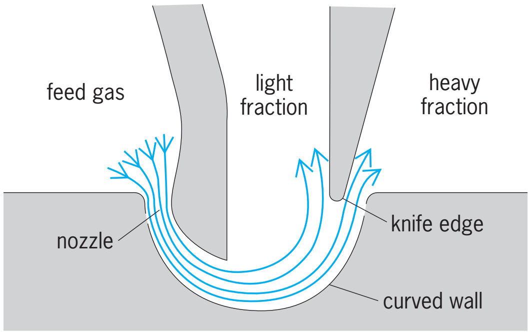


Fig. 4. Cross section of slit-shaped separation nozzle system with streamlines of the gas flow. (*After M. Benedict, T. H. Pigford, and H. W. Levi, Nuclear Chemical Engineering, 2d ed., McGraw-Hill, 1981*)

creates an atomic vapor stream of ^{235}U and ^{238}U. The tuned dye laser beams are passed through the vapor stream, where they photoionize the ^{235}U atoms. An electromagnetic field deflects the selected photoions to the product collector, where they condense. The ^{238}U atoms, which are unaffected by the color-selective laser beams, pass through the product collector to condense on the tails collector. The enriched uranium liquid-metal condensate flows out of the separator to be cast and stored in solid metallic form, or to be converted to uranium hexafluoride or oxide as required.

Nozzle process. In the German jet nozzle or Becker process, which is a laboratory process, uranium hexafluoride gas is diluted in a helium or hydrogen carrier gas, and fed by compressor into the nozzle on one side of a 1–2-m long (3–6-ft) semicircular slot in which there is a deflector blade on the other side to separate enriched and depleted streams (**Fig. 4**). The semicircular wall of the slot provides a partial centrifugal effect such that the heavier isotope tends to stay closer to the wall compared to the lighter isotope. The radius of the semicircular slot is of the order of 0.1 mm and is formed by stacking photoetched foils.

Helikon process. The South African Helikon process is aerodynamically similar to the Becker process from which it was developed. However, it is believed to use a vortex tube instead of a slot and opposing gas flows instead of a deflector blade for flow separation. South Africa's small power-intensive commercial Helikon uranium enrichment plant was shut down in the mid-1990s because of its high cost of operation.

Thermal diffusion. The separation of isotopes by thermal diffusion is based on the fact that when a temperature gradient is established in a mixture of uniform composition, one component will concentrate near the hot region and the other near the cold region. Thermal diffusion is carried out in the annular space between two vertical concentric pipes, the inner one heated and the outer one cooled. Thermal diffusion was used in the United States during World War II for uranium, but it is not efficient for industrial separation. It has been used to separate small quantities of isotopes for research purposes.

Chemical exchange. The chemical exchange process has proved to be the most efficient for separating isotopes of the lighter elements. This process is based on the fact that if equilibrium is established between, for example, a gas and a liquid phase, the composition of the isotopes will be different in the two phases. Thus, if hydrogen gas is brought into equilibrium with water, the ratio of deuterium to hydrogen is several times greater in the water than in the hydrogen gas. By repeating the process in a suitable cascade, it is possible to effect a substantial separation of the isotopes with a relatively small number of stages. The chief use of chemical exchange is in the large-scale production of heavy water. Chemical exchange has also been used for the large-scale separation of other isotopes. For example, the isotopes of boron have been separated by fractional distillation of the boron trifluoride–dimethyl ether complex. *See* CHEMICAL EQUILIBRIUM.

Distillation. The separation of isotopes by distillation is much less efficient than separation by other methods. Distillation was used during World War II to produce heavy water, but the cost was high and the plants are no longer in existence. *See* DISTILLATION.

Electrolysis. Electrolysis of water is the oldest large-scale method of producing heavy water. Under favorable conditions, the ratio of hydrogen to deuterium in the gas leaving a cell in which water is electrolyzed is eight times the ratio of these isotopes in the liquid. Because it is electricity-intensive, the process is very expensive unless low-cost hydroelectric generating capacity is available. Electrolysis is used in the United States only as a finishing step to concentrate heavy water to final-product specifications. *See* ELECTROLYSIS.

Electromagnetic process. Electromagnetic separation was the method which was first used to prove existence of isotopes. The mass spectrometer and mass spectrograph are still widely used as research tools. In the electromagnetic process, vapors of the material to be analyzed are ionized, are accelerated in an electric field, and enter a magnetic field which causes the ions to be bent in a circular path. Since the light ions have less momentum than the heavy ions, they are bent through a circle of smaller radius, and the two isotopes can be separated by placing collectors at the proper location. The large-capacity electromagnetic uranium isotope separators developed in the United States during World War II were called calutrons. *See* CHARGED PARTICLE OPTICS; MASS SPECTROSCOPE.

Julian J. Steyn

Bibliography. M. Benedict, T. H. Pigford, and H. W. Levi, *Nuclear Chemical Engineering*, 2d ed., 1981; Edison Electric Institute, *EEI Enrichment Handbook*, Rep. NFC-90-001, 1990.

Isotope shift

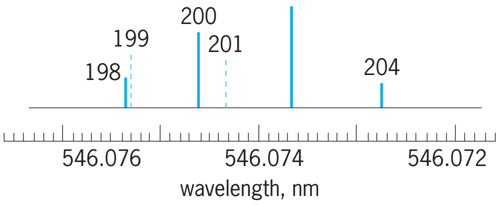
A small difference between the different isotopes of an element in the transition energies corresponding to a given spectral-line transition. For an electronic spectral-line transition between two energy levels *a* and *b* in an atom or ion with atomic number Z, the small difference $\Delta E_{ab} = E_{ab}(A') - E_{ab}(A)$ in the transition energy between isotopes with mass numbers A' and A is the isotope shift. It consists largely of the sum of two contributions, the mass shift (MS) and the field shift (FS), also called the volume shift. The mass shift is customarily divided into a normal mass shift (NMS) and a specific mass shift (SMS); each is proportional to the factor $(A' - A)/A'A$. The normal mass shift is a reduced mass correction that is easily calculated for all transitions. The specific mass shift is produced by the correlated motion of different pairs of atomic electrons and is thus absent in one-electron systems.

It is generally difficult to calculate precisely the specific mass shift, which may be 30 times larger than the normal mass shift for some transitions. The field shift is produced by the change in the finite size and shape of the distribution of nuclear charge (protons) when the number of neutrons $N = A - Z$ varies with Z fixed. Since electrons whose orbits penetrate the nucleus are influenced most, *S-P* and *P-S* transitions have the largest field shift. Measurements of isotope shifts, coupled with accurate theoretical calculations of mass shifts and relativistic and other corrections, can be used to determine the change $\delta\langle r_c^2\rangle$ in the mean-square nuclear charge radius r_c^2 as N varies with Z fixed.

For very light elements, $Z \lesssim 37$, the mass shift dominates the field shift. For $Z = 1$, the 0.13 nanometer shift (0.02%) in the red Balmer line led to the discovery of deuterium, the $A = 2$ isotope of hydrogen. For medium-heavy elements, $38 \lesssim Z \gtrsim 57$, the mass shift and field shift contributions to the isotope shift are comparable. For heavier elements, $Z \gtrsim 58$, the field shift dominates the mass shift. A representative case is shown in the **illustration**. *See* ATOMIC STRUCTURE AND SPECTRA; DEUTERIUM; RYDBERG CONSTANT.

When isotope shift data have been obtained for at least two pairs of isotopes of a given element, a graphical method introduced by W. H. King in 1963 can be used to evaluate quantitatively the separate contributions of the mass shift and the field shift. Experimentally determined field shifts can be used to test theoretical models of nuclear structure, shape, and multipole moments. Experimentally determined specific mass shifts can be used to test detailed theories of atomic structure and relativistic effects. *See* NUCLEAR MOMENTS; NUCLEAR STRUCTURE.

Some atomic transitions exhibit hyperfine structure, which generally is dominated by the interaction of electron orbital and nuclear magnetic dipole moments, the latter existing for isotopes with nonzero nuclear spin. Accurate measurements of magnetic hyperfine structures exhibiting the so-called hyperfine anomaly (Bohr-Weisskopf effect) gives information about how the distribution of nuclear magnetization changes as N varies with Z fixed. *See* HYPERFINE STRUCTURE.



Some isotope shifts in the green line of mercury, $Z = 80$. In this heavy element, the contribution of the field shift is much larger than that of the mass shift.

Experimental techniques that have greatly increased both the amount and the precision of isotope shift data include energetic nuclear reactions for online production of isotopes with half-lives as short as milliseconds, and subsequent high-resolution laser-spectroscopic studies of ions or atoms containing them. Laser spectroscopy of collimated, fast-moving beams of ions or atoms containing the isotopes uses techniques that exploit (rather than are hampered by) the Doppler shift caused by the motion of the ions or atoms through the laser light. Techniques developed for cooling and trapping of neutral atoms or highly charged ions with electromagnetic fields have been extended to short-lived isotopes and allow as few as one trapped atom or ion to be detected. Isotope shifts measured at low Z have tended to focus on precise tests of atomic few-body dynamics and determination of $\langle r_c^2\rangle$ for short-lived, exotic "halo nuclei" such as ^{6}He and ^{11}Li. Isotope shifts measured with laser or x-ray methods at high Z have tended to focus on determination of $\delta\langle r_c^2\rangle$ and tests of atomic structure calculations where many-body, relativistic, quantum-electrodynamical (QED), and other interactions are significant and difficult to treat. These must be well understood for atomic parity-non-conservation (PNC) experiments, which provide stringent and useful tests of electroweak and other theories when different isotopes of high-Z atoms are used. *See* ATOMIC THEORY; DOPPLER EFFECT; ELECTROWEAK INTERACTION; LASER; LASER COOLING; LASER SPECTROSCOPY; PARITY (QUANTUM MECHANICS); PARTICLE TRAP; QUANTUM ELECTRODYNAMICS; RELATIVISTIC QUANTUM THEORY.

Isotope shift data have also been obtained for x-ray transitions of electrons in inner atomic shells and of muons in muonic atoms. *See* LEPTON. Peter M. Koch

Bibliography. J. Bauche and R. J. Campeau, Recent advances in the theory of atomic isotope shift, *Advan. At. Mol. Phys.*, 12:39–86, 1976; J. C. Berengut, V. V. Flambaum, and M. G. Kozlov, Calculation of relativistic and isotope shifts in Mg I, *Phys. Rev. A*, 72:044501, 2005; S. R. Elliott, P. Beiersdorfer, and M. H. Chen, Trapped-ion technique for measuring the nuclear charge radii of highly charged radioactive isotopes, *Phys. Rev. Lett.*, 76:1031–1034, 1995; E. Gomez, L. A. Orozco, and G. D. Sprouse, Spectroscopy with trapped francium: Advances and perspectives for weak interaction studies, *Rep. Prog. Phys.*, 69:79–118, 2006; R. Sánchez et al., Nuclear charge radii of ^{9,11}Li: The influence of halo electrons, *Phys. Rev. Lett.*, 96:033002, 2006; L.-B. Wang et al., Laser spectroscopic determination of the ^{6}He nuclear charge radius, *Phys. Rev. Lett.*, 93:142501, 2004.

Isotopic irradiation

The uses of the radiation emitted by radioactive isotopes (radioisotopes), principally in industry and medicine. Although this article deals primarily with gamma irradiation, many other forms of radiation from many different sources are often used for irradiation purposes. The use of radioactive isotopes to irradiate various materials can be divided into a number of major areas, including industrial, medical, and research (**Table 1**). For most applications the radioactive isotope is placed inside a capsule, but for some medical applications it is dispersed in the material to be irradiated. *See* RADIOISOTOPE.

The radiation from radioisotopes produces essentially the same effect as the radiation from electron linear accelerators and other high-voltage particle accelerators, and the choice of which to use is based primarily on convenience and cost. Although the radioisotope radiation source does not require the extensive and complex circuitry necessary for a high-voltage radiation source, its radiation is always present and requires elaborate shielding for health protection and specialized mechanisms for bringing the irradiated objects into and out of the radiation beam. Further, the radiation output decreases with time according to the half-life of the radioisotope, which must therefore be replaced periodically. *See* PARTICLE ACCELERATOR.

Industrial applications. The two main radioisotopes for industrial processing are cobalt-60 with a half-life of 5.271 years and an average gamma-ray energy of 1.25 MeV, and cesium-137 with a half-life of 30.07 years and a gamma-ray energy of 0.662 MeV. *See* CESIUM; COBALT.

The application of industrial irradiation is increasing, with sterilization of medical disposables using cobalt-60 gamma rays being the most common. This includes the irradiation of plastic syringes, surgical gloves, and nonwoven drapes and gowns. In addition to the medical supplies, gamma sterilization of disposable laboratory supplies—petri dishes, pipettes, and test tubes—is also a fast-growing application.

A promising and appealing application is food preservation, including the reduction of postharvest losses by the elimination of pests with irradiation. Other food irradiation effects can be accomplished at distinct irradiation doses (**Table 2**). Regulatory bodies in several countries approve the consumption of over 20 irradiated foods. Extensive research on this subject has been done at the Bhabha Atomic Research Center in Trombay, India, where the need for bulk sterilization of food is more pressing.

TABLE 2. Food irradiation effects

Effect	Irradiation dose
Inhibition of sprouting	6–15 krad (60–150 gray)
Disinfection	20–1000 krad (200–1000 Gy)
Control of parasites	20–200 krad (200–2000 Gy)
Shelf-life extension	200–500 krad (2000–5000 Gy)
Extended refrigerated storage	200–500 krad (2000–5000 Gy)
Elimination of pathogens	300–700 krad (3000–7000 Gy)
Reduction of microbial load	Up to 1 Mrad (10 kGy)
Sterilization	Over 1 Mrad (10 kGy)

The biocidal effect of the gamma-irradiation process is effective for the control of microbiological contamination in many raw materials used by the pharmaceutical and cosmetic industries. Such materials include distilled water, starch powder, rice powder, lactose, talc, cocoa butter, inorganic pigments, petroleum jelly, and enzymes. Other applications include irradiation of male insects for pest control, sterilization of corks, sterilization of laboratory animal bedding, preparation of vaccines, degradation of poly(tetrafluoroethylene) scraps used in lubricants, cross-linking of polyethylene for shrink films, and production of wood-polymer and concrete-polymer composites.

Cesium-137 is also used for sludge processing to produce fertilizer or cattle feed. It is available in large quantities in the reprocessing of fuel elements from nuclear reactors, and the process helps solve the problem of disposing of cesium from this source. *See* NUCLEAR FUELS REPROCESSING.

Medical applications. Radioisotopes are used in the treatment of cancer by radiation. Encapsulated sources are used in two ways: the radioisotopes may be external to the body and the radiation allowed to impinge upon and pass through the patient (teletherapy), or the radiation sources may be placed within the body (brachytherapy).

For teletherapy purposes, cobalt-60 is the most commonly used isotope, with source strengths ranging from several thousand curies up to 10^4 Ci (3.7×10^{14} becquerels). Some cesium-137 irradiators have been built, but cesium-137 radiation is not as penetrating as that from cobalt-60.

TABLE 1. Most commonly used radioisotopes and major applications

Isotope	Half-life	Average γ-ray energy, MeV	Main uses
Radium-226 in equilibrium (^{226}Ra)	1600 years	0.83	Medical (brachytherapy)
Gold-198 (^{198}Au)	2.695 days	0.412	Medical (brachytherapy—permanent implants)
Iridium-192 (^{192}Ir)	73.8 days	0.38	Medical (brachytherapy)
Cesium-137 (^{137}Cs)	30.07 years	0.662	Industrial, medical (brachytherapy), research irradiators
Cobalt-60 (^{60}Co)	5.271 years	1.25	Industrial, medical (teletherapy), research irradiators
Iodine-125 (^{125}I)	59.4 days	0.035	Medical (brachytherapy—permanent implants)
Iodine-131 (^{131}I)	8.021 days	0.187, average β-ray energy	Medical (dispersal technique)

Historically, radium-226 has been used as the brachytherapy encapsulated source, but due to serious problems that can arise from a health physics aspect if the capsule breaks, radium is being replaced with cesium-137 sources. Iridium-172 is also used for brachytherapy applications. *See* HEALTH PHYSICS.

In order to avoid the long half-life of radium, its daughter product radon-222, which is an inert gas, was used in very small glass or gold seeds. Today radon has been replaced by iodine-125 or gold-198 seeds, which are put directly in a tumor and permanently left in place. *See* RADIUM.

For some medical applications the radioisotope is dispersed in the body; the most commonly used is iodine-131. The specific advantage of internal therapy with iodine-131 is that the thyroid gland concentrates the element iodine. When the radioisotope is administered either orally or intravenously in a highly purified form, it goes to the thyroid, where certain forms of thyroid disorders and cancers can be treated by the radiation. Unlike the other radioisotopes discussed here, the iodine-131's therapeutic effectiveness depends upon the beta rays emitted, not the gamma rays. *See* IODINE.

Another medical use for radioisotopes is the requirement that all transfusions of blood be irradiated before being given to certain patients so that most of the lymphocytes present are destroyed. For this purpose, self-contained irradiators are used. Cobalt-60 or cesium-137 encapsulated sources are permanently fixed and sealed off within the radiation shield, and the irradiation chamber is introduced into the radiation field by rotation. Such units are also extensively used for radiobiological and other research purposes. *See* RADIOACTIVITY AND RADIATION APPLICATIONS; RADIOLOGY. Peter R. Almond

Bibliography. D. A. Bromley, Neutrons in science and technology, *Phys. Today*, 36(12):31–39, December 1983; Institute of Food Technologists, Food irradiation (a collection of articles), *Food Technol.*, 27(2):38–60, February 1983; T. Ouwerkerk, An overview of the most promising industrial applications of gamma processing, *International Symposium on Application and Technology of Ionizing Radiations*, Riyadh, Saudi Arabia, March 1982; M. Satin, *Food Irradiation: A Guidebook*, 2d ed., 1996.

Ixodides

A suborder of the Acarina, class Arachnida, comprising the ticks. Ticks differ from mites, their nearest relatives, in their larger size and in having a pair of breathing pores, or spiracles, behind the third or fourth pair of legs. They have a gnathosoma (or so-called head or capitulum), which consists of a base (basis capituli), a pair of palps, and a rigid, elongated, ventrally toothed hypostome which anchors the parasite to its host. They also have a pair of protrusible cutting organs, or chelicerae, which permit the insertion of the hypostome. The stages in the life cycle are egg, larva, nymph, and adult. The larvae have three pairs of legs; nymphs and adults, four. The 600 or so known species are all bloodsucking, external parasites of vertebrates including amphibians, reptiles, birds, and mammals.

Ticks are divided into three families, Argasidae, Ixodidae, and Nuttalliellidae. The last contains but one exceedingly rare African species, *Nuttalliella namaqua*, morphologically intermediate between the Argasidae and the Ixodidae. It is of no known importance, either medically or economically.

Argasidae. Argasids, or the soft ticks, differ greatly from ixodids in that the sexes are similar; the integument of adults and nymphs is leathery and wrinkled; there is no dorsal plate, or scutum; the gnathosoma is ventral in adults and nymphs but anterior in larvae; and the spiracles are small and anterior to the hindlegs (**Fig. 1**). These ticks frequent nests, dens, and resting places of their hosts. Adults feed intermittently and eggs are laid a few at a time in niches where the females seek shelter. Larvae feed for a few minutes to several days, then detach and transform to nymphs which feed and molt several times before transforming to adults. Nymphs and adults are notably resistant to starvation; some are known to live 10 years or longer without feeding.

The family contains about 85 species, with 20 in the genus *Argas* and 60 in *Ornithodoros*. Several are of medical or veterinary importance. *Argus persicus*, *A. miniatus*, *A. radiatus*, *A. sanchezi*, and *A. neghmei* are serious pests of poultry. Some of these species carry fowl spirochetosis, a disease with a high mortality rate. Larvae and nymphs of *Otobius megnini*, the spinose ear tick, feed deep in the ears of domesticated animals in many semiarid regions. Heavy infestations cause intense irritation which leads to unthriftiness and sometimes to death. The life history is unusual in that the adults do not feed. *Ornithodoros moubata* transmits

Fig. 1. Argasid tick, *Ornithodoros coriaceus*, shown enlarged.

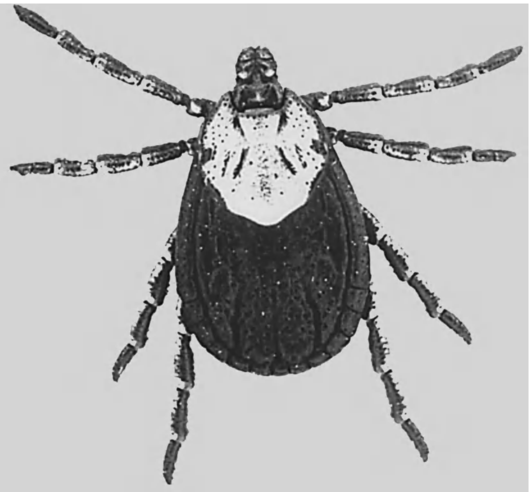


Fig. 2. Ixodid tick, ***Dermacentor andersoni***, female, shown enlarged.

relapsing fever in East, Central, and South Africa. It is a highly domestic parasite and humans are probably the chief host. *Ornithodoros turicata*, *O. hermsi*, *O. talaje*, and *O. rudis* are important vectors of relapsing fever in the Western Hemisphere, as is *O. tholozani* in Asia. The bites of most species that attack humans produce local and systemic reactions; the bites of some species, especially *O. coriaceus* of California and Mexico, are extremely venomous. *See* RELAPSING FEVER.

Ixodidae. In contrast to argasids, Ixodidae have a scutum covering most of the dorsal surface of the male but only the anterior portion of females, nymphs, and larvae (**Fig. 2**). They are known as the hard ticks. The sexes are thus markedly dissimilar. The gnathosoma extends anteriorly, and the large spiracles are posterior to the hindlegs. Instead of frequenting nesting places, these ticks are usually more or less randomly distributed throughout their hosts' environment. Larvae, nymphs, and adults feed but once, and several days are required for engorgement. The immature stages of most species drop to the ground for molting but those of the genus *Boophilus* and a few others molt on the host. The female lays a mass containing up to 10,000 or more eggs. The life cycle is usually completed in 1–2 years.

The family consists of about 11 well-defined genera with 500 species. Many species transmit disease agents to humans and animals; included are viruses, rickettsiae, bacteria, protozoa, and toxins. Transmission is by bite or by contact with crushed tick tissues or excrement. Virus diseases of humans include Colorado tick fever of the western United States, transmitted by *Dermacentor andersoni*, and Russian spring-summer encephalitis and related diseases in Europe and Asia, transmitted by *Ixodes persulcatus*. Some important rickettsial diseases of humans are spotted fever, widely distributed in the Western Hemisphere, of which *D. andersoni*, *D. variabilis*, *Rhipicephalus sanguineus*, and *Amblyomma cajennense* are important vectors; boutonneuse fever and related diseases of the Mediterranean region and Africa transmitted by *R. sanguineus* and some other species; and Q fever, the agent of which, although occurring in many ticks throughout the world, is not commonly transmitted directly by them. Tularemia (rabbit fever) is a bacterial disease transmitted by several species of ticks in the Northern Hemisphere. Tick paralysis, which is believed to be caused by toxins, occurs in many parts of the world and is produced by several species of ticks during the feeding process.

The numerous tick-borne diseases of animals cause vast economic losses, especially in tropical and subtropical regions. Examples are babesiasis, a protozoan disease caused by species of *Babesia* including *B. bigemina*, the agent of the widely distributed Texas cattle fever, which is transmitted principally by species of *Boophilus*; East Coast fever of Africa, another protozoan disease, caused by *Theileria parva*, is carried by several species of *Rhipicephalus*. Aside from carrying disease, several species are extremely important pests of humans and animals. Heavy infestations of ticks produce severe anemia in livestock, and even death, from the loss of blood alone. *See* TICK PARALYSIS; TULAREMIA.

Glen M. Kohls

J

J/psi particle — Jute

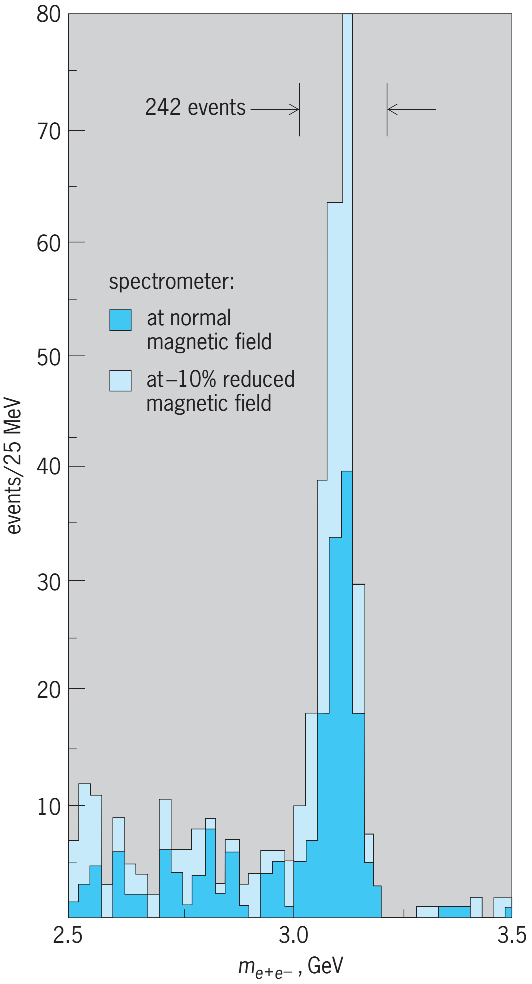
J/psi particle

An elementary particle with an unusually long lifetime or, from the Heisenberg uncertainty principle, with an extremely narrow width $\Gamma = 86.6 \pm 6.0$ keV, and a large mass $m = 3096.93 \pm 0.09$ MeV. It is a bound state containing a charm quark and an anticharm quark. The discovery of the J/psi particle is one of the cornerstones of the standard model.

Discovery. Prior to 1974, there was no theoretical justification, and no predictions existed, for long-lived particles in the mass region 1–10 GeV/c^2.

The J/psi particles are rarely produced in proton-proton (p-p) collisions. Statistically, they occur once after hundreds of millions of subnuclear reactions, in which most of the particles are "ordinary" elementary particles, such as the kaon (K), pion (π), or proton (p). Searches for the J/psi particle are conducted by detecting its electron-positron (e^+e^-) decays. A two-particle spectrometer was used to discover this particle. A successful experiment must have: (1) a very high intensity incident proton beam to produce a sufficient amount of J particles for detection; and (2) the ability, in a billionth of a second, to pick out the J/psi → e^-e^+ pairs amidst billions of other particles through the detection apparatus. *See* PARTICLE ACCELERATOR.

The detector is called a magnetic spectrometer. A positive particle and a negative particle each traversed one of two 70-ft-long (21-m) arms of the spectrometer. The e^+ and e^- were identified by the fact that a special counter, called a Cerenkov counter, measured their speed as being slightly greater than that of all other charged particles. Precisely measured magnetic fields bent them and measured their energy. Finally, as a redundant check, the particles plowed into high-intensity lead-glass and the e^+ and e^- immediately transformed their energy into light. When collected, this light "tagged" these particles as the e^+ and e^-, and not heavier particles such as the π, K, or p. The simultaneous arrival of an e^- and



On-line data from August and October 1974, showing existence of J/psi particle. (*After J. J. Aubert et al., Discovery of the new particle*, J. Nuc. Phys., *B89*(*1*):*1–18*, *1975*)

e^+ in the two arms indicated the creation of high-energy light quanta from nuclear interactions. The sudden increase in the number of e^+e^- pairs at a given energy (or mass) indicated the existence of a new particle (see **illus.**). *See* CERENKOV RADIATION; PARTICLE DETECTOR.

Properties. Since its discovery in 1974, more than 10^9 J/psi particles have been produced. More than 100 different decay modes and new particles radiating from the J/psi particle have been observed. The J/psi particle has been shown to be a bound state of charm quarks. The long lifetime of the J/psi results from its mass being less than the masses of particles which separately contain a charm and an anticharm quark. This situation permits the J/psi to decay only into noncharm quarks, and empirically this restriction was found to lead to a suppression of the decay rate resulting in a long lifetime and narrow width. The subsequent discovery of the *b* quark and the intermediate vector bosons Z^0 and $W^{\pm}$, and studies of Z^0 decays into charm, *b*, and other quarks, show that the theory of the standard model is in complete agreement with experimental data to an accuracy of better than 1%. *See* CHARM; ELEMENTARY PARTICLE; INTERMEDIATE VECTOR BOSON; QUARKS; STANDARD MODEL; UPSILON PARTICLES.

Samuel C. C. Ting

Bibliography. O. Adriani et al., Results from the L3 experiment, *Phys. Rep.,* 236:1–146, 1993; J. J. Aubert et al., *Phys. Rev. Lett.*, 33:1404 and 1624, 1974; E. J. Augustin et al., *Phys. Rev. Lett.*, 33:1406, 1974; W. Braunschweig et al., *Phys. Lett.*, 53B:393, 1974; K. Hikasa et al., Review of particle properties, *Phys. Rev. D*, vol. 45, no. 11, pt. 2, 1992.

Jade

A name that may be applied correctly to two distinct minerals. The two true jades are jadeite and nephrite. In addition, a variety of other minerals are incorrectly called jade. Idocrase is called California jade, dyed calcite is called Mexican jade, and green grossularite garnet is called Transvaal or South African jade. The most commonly encountered jade substitute is the mineral serpentine. It is often called new jade or Korean jade. The most widely distributed and earliest known true type is nephrite, the less valuable of the two. Jadeite, the most precious of gemstones to the Chinese, is much rarer and more expensive.

Nephrite. Nephrite is one of the amphibole group of rock-forming minerals, and it occurs as a variety of a combination of the minerals tremolite and actinolite. Tremolite is a calcium-magnesium-aluminum silicate, whereas iron replaces the magnesium in actinolite. Although single crystals of the amphiboles are fragile because of two directions of easy cleavage, the minutely fibrous structure of nephrite makes it exceedingly durable. It occurs in a variety of colors, mostly of low intensity, including medium and dark green, yellow, white, black, and blue-gray. Nephrite has a hardness of 6 to $6^1/_2$ on Mohs scale, a specific gravity near 2.95, and refractive indices of 1.61 to 1.64. On the refractometer, nephrite gemstones show a single index near 1.61. Nephrite occurs throughout the world; important sources include Russia, New Zealand, Alaska, several provinces of China, and a number of states in the western United States. *See* AMPHIBOLE; GEM; TREMOLITE.

Jadeite. Jadeite is the more cherished of the two jade minerals, because of the more intense colors it displays. It is best known in the lovely intense green color resembling that of emerald (caused by a small amount of chromic oxide). In the quality known as imperial jade, the material is at least semitransparent. White, green and white, light reddish violet, bluish violet, brown, and orange colors are also found. Jadeite also has two directions of easy cleavage, but a comparable random fibrous structure creates an exceptional toughness. Although jadeite has been found in California and Guatemala, the only important source of gem-quality material ever discovered is the Mogaung region of upper Burma. The hardness of jadeite is $6^1/_2$ to 7, its specific gravity is approximately 3.34, and its refractive indices are 1.66 to 1.68; on a refractometer, only the value 1.66 is usually seen. *See* JADEITE.

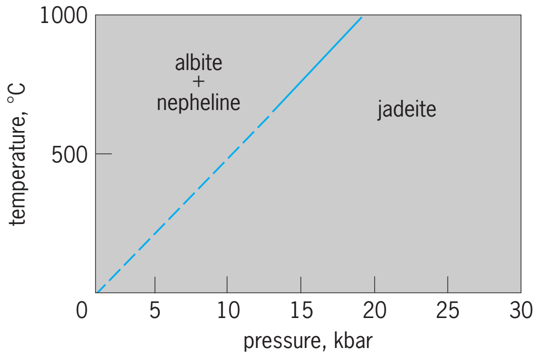
Richard T. Liddicoat, Jr.

Jadeite

The monoclinic sodium aluminum pyroxene, $NaAlSi_2O_6$. Free crystals are rare. Jadeite usually occurs as dense, felted masses of elongated blades or as fine-grained granular aggregates. It has a Mohs hardness of 6.5 and a density of 3.25–3.35. It has a vitreous or waxy luster, and is commonly green but may also be white, violet, or brown. Jadeite exhibits the 93° and 87° pyroxene cleavages and a splintery fracture. It is extremely tough.

Jadeite is always found in metamorphic rocks. It is associated with serpentine at Tawmaw, Burma; Kotaki, Japan; and San Benito County, California. It occurs in metasedimentary rocks of the Franciscan group in California and in Celebes. It is also found in Tibet; Yunan Province, China; and Guatemala. In addition, the jadeite “molecule” is present in high concentrations in solid solution in omphacite, a dark green pyroxene which, like garnet, is a major constituent of eclogite, a high-pressure, high-temperature metamorphic rock formed in the deep crust or upper mantle. Eclogites are found in the Bessi district, Japan; Mampong, Ghana; Gertrusk, Austria; Sonoma County, California; and as nodules in South African kimberlite pipes. *See* ECLOGITE; SOLID SOLUTION.

Jadeite’s composition is intermediate between that of nepheline and albite, but jadeite does not crystallize in the binary system $NaAlSiO_4$–$NaAlSi_3O_8$, or in the ternary system Na_2O–Al_2O_3–SiO_2, at low pressures. Its high density compared with the above tektosilicates and its presence in eclogite suggest that it is a high-pressure phase. This was confirmed by high pressure experiments on stability relationships between jadeite, nepheline, and albite (see **illus.**). Pure jadeite usually occurs with minerals indicative



Equilibrium curve for the reaction albite + nepheline → 2 jadeite. 1 kilobar = 10 pascals. °F = (°C × 1.8) + 32.

of relatively low temperatures (390–570°F or 200–300°C) such as analcite, lawsonite, chlorite, glaucophane, stilpnomelane, and zeolites.

Jadeite is valued as a precious stone for carvings. *See* HIGH-PRESSURE MINERAL SYNTHESIS; JADE; PYROXENE. Lawrence Grossman

Jahn-Teller effect

A distortion of a highly symmetrical molecule, which reduces its symmetry and lowers its energy. The effect occurs for all nonlinear molecules in degenerate electronic states, the degeneracy of the state being removed by the effect. It was first predicted in 1937 by H. A. Jahn and E. Teller. In early experimental work, the effect often "disappeared" or was masked by other molecular interactions. This has surrounded the Jahn-Teller effect with a certain mystery and allure, rarely found in science today. However, there are now a number of clear-cut experimental examples which correlate well with theoretical predictions. These examples range from the excited states of the most simple polyatomic molecule, H_3, through moderate-sized organic molecules, like the ions of substituted benzene, to complex solid state phenomena involving crystals or localized impurity centers. *See* DEGENERACY (QUANTUM MECHANICS); MOLECULAR STRUCTURE AND SPECTRA; QUANTUM MECHANICS.

Square planar molecule example. The fact that electronic degeneracy may destroy the symmetry upon which it is based is most easily demonstrated for molecules with fourfold symmetry. Consider a square planar molecule (with fourfold symmetry, D_{4h}) with an unpaired electron in an orbitally degenerate state (for example, E_u). This state will be characterized by two orthogonal wavefunctions ψ_A and ψ_B and electron density distributions with mutually perpendicular nodal planes, as shown schematically in **Fig. 1*a*** and *b*. As long as the nuclei are undisplaced, the molecules in Fig. 1*a* and 1*b* are geometrically congruent and the energy of the system will be the same, regardless of whether the electron is in orbital ψ_A or ψ_B, that is, $E_A(0) = E_B(0)$. Imagine the molecule undergoing an in-plane bending vibration ($B_{2g}\nu_4$) and consider the nuclei displaced an increment $+q$ along the appropriate normal coordinate Q. It is easily seen that the positions of the nuclei with respect to the electronic distributions characterized by ψ_A and ψ_B will not be equivalent, and the energies of the states A and B will differ, as shown on the right side of the potential energy diagram, Fig. 1*c*. The same argument will, of course, hold for the opposite phase of the vibration and displacement $-q$ as shown on the left side of Fig. 1*c*. It is easily seen that the molecule in Fig. 1*a* with displacement $+q$ is geometrically congruent with the molecule in Fig. 1*b* for displacement $-q$. It can therefore be deduced that for the electronic energies $E_A(q) = E_B(-q)$ and similarly $E_A(-q) = E_B(q)$. The establishment of these five points determines the general form of the Jahn-Teller potential curve in Fig. 1*c*. The potential surfaces of the states A and B must cross. The crossing point at zero displacement, with the corresponding energy $E(0)$, is clearly not an absolute minimum; therefore, the totally symmetric square configuration will not be a stable equilibrium for the degenerate electronic state. Therefore, at equilibrium, the molecule will be distorted from a square, and its energy will be lowered. *See* MOLECULAR ORBITAL THEORY.

The above arguments are not restricted to square molecules. With the exception of linear molecules which suffer Renner-Teller effects, all polyatomic molecules of sufficiently high symmetry to possess

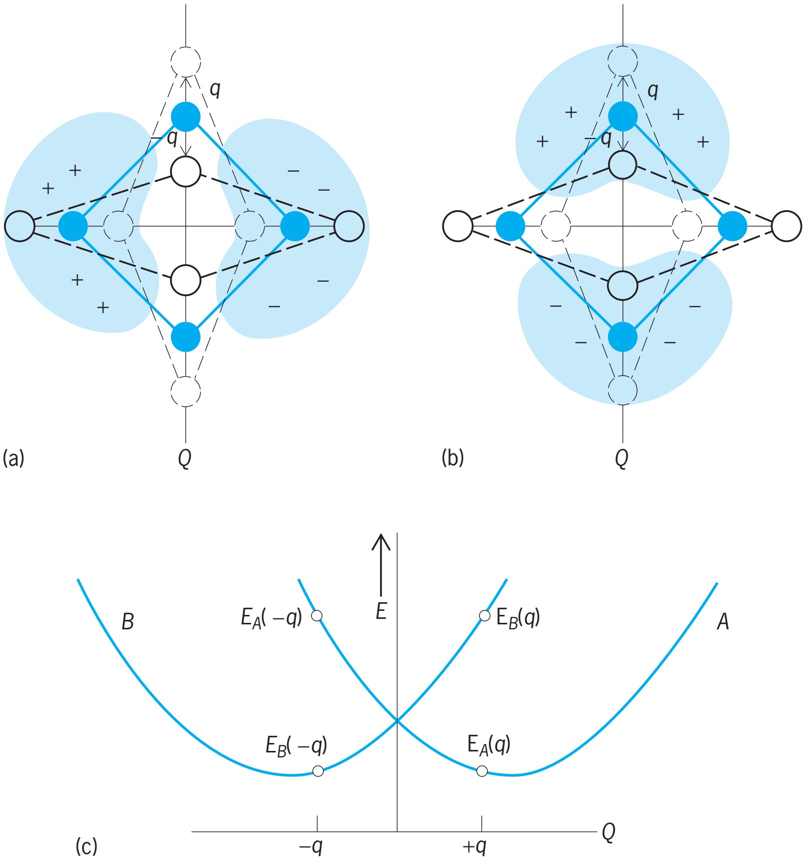


Fig. 1. Jahn-Teller effect for a square planar molecule. Displacement of the nuclei of the molecule in the electronic field provided by (*a*) the wavefunction Ψ_A and (*b*) Ψ_B. (*c*) Energy *E* corresponding to Ψ_A and Ψ_B as a function of the displacement coordinate *Q*.

orbitally degenerate electronic states will be subject to the Jahn-Teller instability. However, in cases other than the fourfold symmetry discussed above, the proof is somewhat more involved and requires the use of the principles of group theory. *See* GROUP THEORY; RENNER-TELLER EFFECT.

Vibrational structure. The distortion of the electronic potential surface of a degenerate electronic state by the Jahn-Teller effect has been considered. The nuclear motion on such a distorted potential surface leads to perturbations in the vibrational structure. These perturbations result in shifts and splittings of the vibrational levels. While group theory, as noted above, can demonstrate the susceptibility of a given molecule to Jahn-Teller instability, it gives no information about the magnitude of the effect; such information must come from experiments. The hexagonal benzene molecule, C_6H_6, and its symmetrically substituted derivatives, represent almost ideal examples of highly symmetrical molecules. Their least tightly bound electrons reside in a pair of degenerate π orbitals. The removal of one of these electrons produces a benzene positive ion whose ground state is electronically degenerate and so fulfills the criteria for a Jahn-Teller effect. These ions have been studied extensively experimentally and represent excellent examples with which to illustrate vibrational effects.

For molecules with the sixfold symmetry (D_{6h}) of a regular hexagon, group theoretical arguments demonstrate that Jahn-Teller distortions of the nuclei occur for motions along four doubly degenerate e_{2g} vibrational normal modes denoted ν_{15}–ν_{18}. These are shown schematically in **Fig. 2**, where the arrows represent the cartesian components of the atomic displacements. The three-dimensional form of the potential energy E with respect to any one of these doubly degenerate coordinates Q_i is shown in **Fig. 3***a*; Fig. 3*b* shows cuts through this type of surface assuming various magnitudes for the Jahn-Teller distortion and illustrates how the introduction of the Jahn-Teller effect changes the potential surface. On the far left with no Jahn-Teller effect, the doubly degenerate electronic potential has a minimum at the symmetrical configuration (a regular hexagon). As the Jahn-Teller effect increases, the electronic potentials split and lowered minima (Jahn-Teller stabilization) in the potential occur at distorted nuclear configurations.

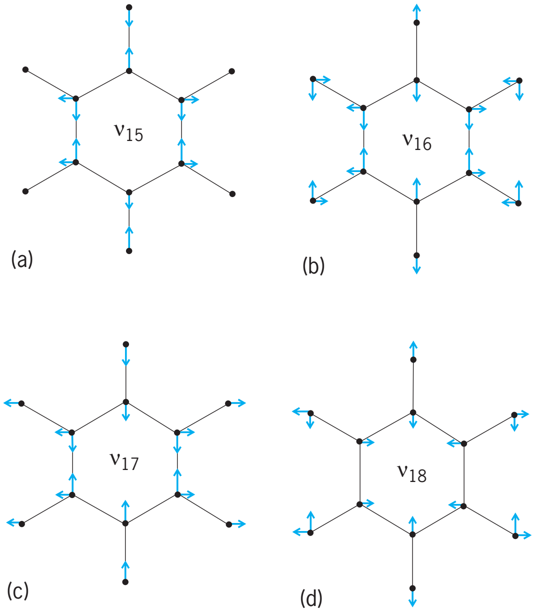


Fig. 2. The four e_{2g} vibrational modes which are Jahn-Teller-active for the positive ion of any substituted benzene with D_{6h} symmetry: (*a*) C-C stretch, (*b*) C-F stretch, (*c*) C-C-C bend, and (*d*) C-F bend.

The Jahn-Teller effect is usually characterized in terms of a linear parameter D defined by $D = E_{JT}/\hbar\omega$, where E_{JT} is the Jahn-Teller stabilization energy (the separation between the central maximum and the distorted minima) and $\hbar\omega$ is the energy of the vibrational mode leading to the instability. As long as only the linear Jahn-Teller effect (wherein only terms linear in the displacement coordinate appear in the hamiltonian) is considered, the distorted potential of the D_{6h} species in Fig. 3*a* exhibits an equipotential moat around the central maximum. The quadratic Jahn-Teller effect (caused by terms quadratic in the displacement coordinate) produces the threefold local minima and saddle points illustrated in Fig. 3*a*, and obviously requires the introduction of an additional coupling parameter, analogous to D, for its characterization. For small distortion when D is much less than 1 and quadratic terms are insignificant, approximate formulas describing the vibronic levels have been derived. For the general case when the distortion is large (D approximately 1 or greater) or when terms quadratic in Q are important, no explicit formulas can usually be given. More sophisticated techniques are needed to calculate the vibrational structure, and usually involve the diagonalization of large matrices.

Such analysis shows that for the substituted benzenoid cations mentioned above the linear Jahn-Teller parameter D is not greater than about 1.0, while quadratic and higher terms are almost negligible. The energy reduction, that is, the Jahn-Teller stabilization energy E_{JT}, is about 2–3 kcal/mole (8–12 kJ/mole) for these ions. This energy reduction is obtained with the benzene ring distorted from a regular hexagon by 2–3° in its internal angles, and with the C-C bond lengths differing by about 4 picometers.

Static and dynamic effects. The terms static and dynamic Jahn-Teller effect are frequently used, but unfortunately their use is not standardized. The term static Jahn-Teller effect is often used to describe the distortion of the electronic potential as seen in Figs. 1*c* and 3, with corresponding use of dynamic Jahn-Teller effect to describe vibrational motion on this surface. However, other uses of the terms static and dynamic involve the magnitude of the Jahn-Teller effect and the time scale of the experiments used to measure its properties. As long as the distortion is small (D less than 1), all vibrational levels ν will be located above the central maximum (Fig. 3*b*). On the time scale of most experiments, a time-averaged, symmetrical configuration will be sampled, and the term dynamic Jahn-Teller effect is therefore often applied. For a sufficiently large distortion (*D* greater than 1), one or more vibrational levels may be localized in separate, disjoint minima, with interconversions between them being slow. This situation (*D* greater than 1) correspondingly would be referred to as the static Jahn-Teller effect. The distortion will therefore manifest itself as a static or dynamic effect depending on the time scale of the experiment. Further complication arises when the Jahn-Teller active mode is itself degenerate, as in the case of the e_{2g} vibrations of Fig. 2. In this case the switching between the potentials U' and U' of Fig. 3 can occur by pseudorotation within the moat, without ever surmounting the symmetric maximum. Then large quadratic terms, resulting in deep minima around the moat, are required to produce a really static effect.

Solids. While the Jahn-Teller effect was originally postulated for highly symmetric discrete molecules, it can also appear in crystals and other highly symmetric extended systems. Searches for the observable consequences of the Jahn-Teller effect in molecular spectra were initially hampered by the fact that typically only excited electronic states of stable molecules or states of transient species—molecular ions and free radicals—possess the orbital degeneracy required for its observation. Most of the early experimental observations therefore involved optical or microwave spectra of point defects and impurity centers in solids. *See* CRYSTAL DEFECTS.

In treating these systems, usually only the impurity or defect and its nearest-neighbor atoms are considered, and the rest of the solid is regarded as a featureless heat bath. As an example, **Fig. 4*a*** shows schematically a simple cluster, consisting of an electron in a *p* state trapped at a negative ion vacancy in an alkali halide (a so-called F-center), surrounded by a regular octahedron of its nearest neighbor positive ions. In the undistorted lattice, the three possible states, p_x, p_y, and p_z, will be obviously degenerate. Now suppose the cluster undergoes an asymmetric vibration and the ions start moving in the directions indicated by the arrow. It is easily seen that the energy of the p_z state will be initially lowered, since the positive ions will approach more closely the negative lobes of the p_z wave function, while, by a similar argument, those of the p_y and p_x states will be raised (Fig. 4*b*). The surrounding crystal will counterbalance the distortion, with the corresponding energy increase being, by Hooke's law, initially quadratic in displacement. As in the molecular case, it is easy to see that an energy minimum for the F-center electron will not be in the symmetric configuration but at some finite displacement. *See* COLOR CENTERS.

Besides the optical and microwave spectra, the consequences of the distorted geometry are also observable in electron spin resonance and Zeeman spectra. When, for instance, a magnetic field *H* is applied parallel to the *z* axis of Fig. 4*a*, the degeneracy of the *p* state is split into three states with energies of approximately $-\beta H$, corresponding to the three possible angular momentum projection eigenvalues of $L_x = +1, 0, -1$, with β the Bohr magneton. In the Jahn-Teller distorted case, the electron can no longer rotate freely, it has to "drag" its distortion with it. This increases its effective mass and quenches its orbital contribution to the magnetic splitting. Such reductions have been observed in the Zeeman spectra of numerous transition-metal ions in crystals. The magnitude of this reduction is dependent on the Jahn-Teller stabilization energy. At absolute zero (0 K) it was shown to be approximately exp $(-3E_{JT}/2\hbar\omega)$. *See* ELECTRON PARAMAGNETIC RESONANCE (EPR) SPECTROSCOPY; ZEEMAN EFFECT.

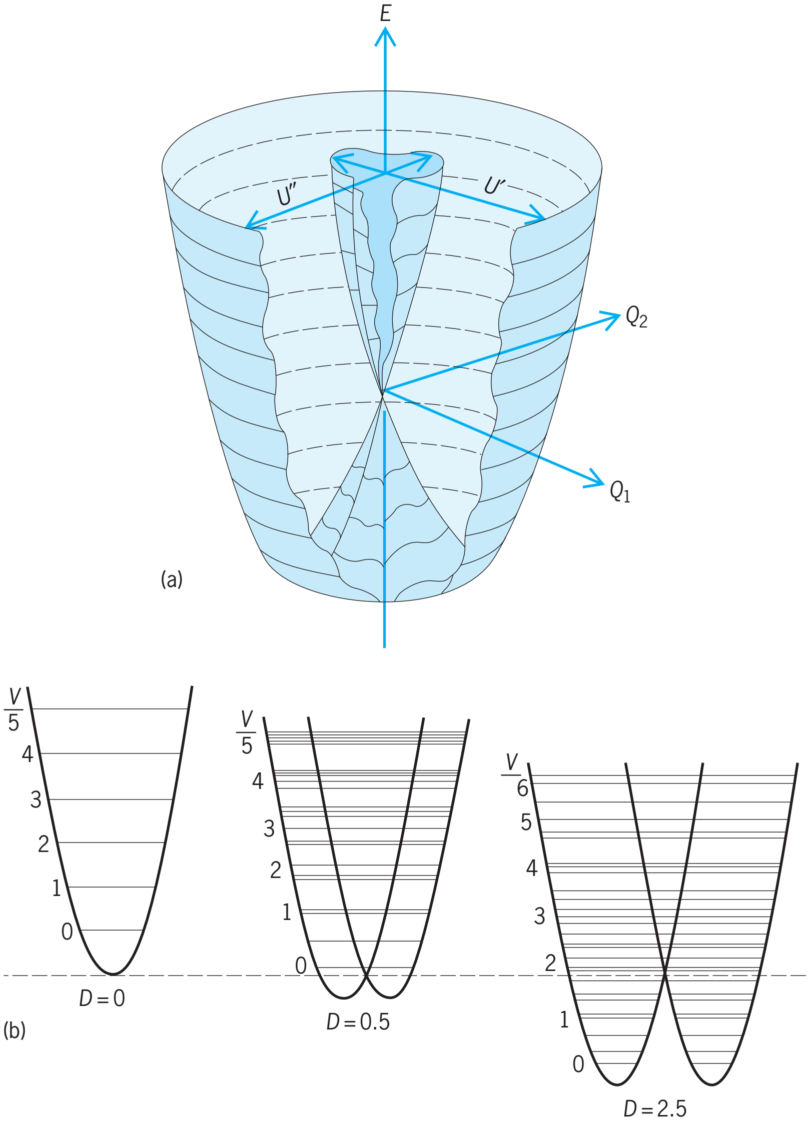


Fig. 3. Potential energy for a doubly degenerate, Jahn-Teller active vibration of a molecule with D_{6h} or D_{3h} symmetry. (*a*) Potential energy surface. (*b*) Cuts through the surface for increasing magnitude of the Jahn-Teller effect going from left to right. Horizontal lines represent vibrational levels, which are numbered with the index *V* for the undistorted molecule.

In some ions the electron-lattice interaction is so weak that even though their ground states are orbitally degenerate, Jahn-Teller effects are ordinarily not detected when they result from isolated impurities. In crystals containing large concentrations

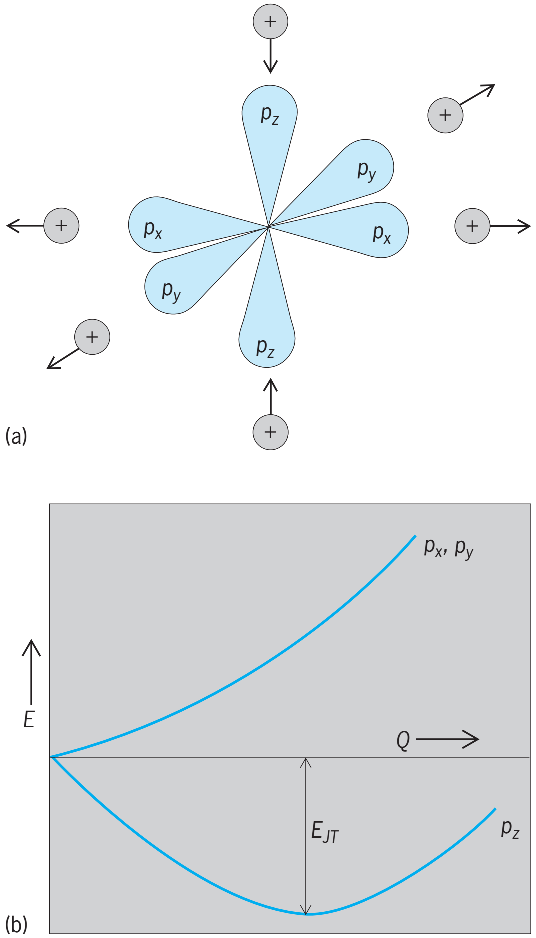


Fig. 4. Jahn-Teller effect in an F-center. (*a*) The *p* orbitals of the F-center surrounded by an octahedron of positive ions. (*b*) The potential energy *E* of the *p* orbitals as the positive ions are displaced, as a function of displacement coordinate *Q*.

of these ions, the ion-ion interactions can greatly enhance these effects. Many rare-earth compounds have been found to undergo second-order phase transitions in which the overall crystal symmetry is reduced and the electronic degeneracy removed in a manner expected for the Jahn-Teller effect. Such distortions, involving cooperation of many ions, are usually called the cooperative Jahn-Teller effect. These minute distortions may cause shifts in phonon modes, which are observable in Raman spectra or neutron diffraction experiments. *See* NEUTRON DIFFRACTION; RAMAN EFFECT.

V. E. Bondybey; Terry A. Miller

Bibliography. I. B. Bersuker, *The Jahn-Teller Effect*, 2006; C. C. Chancey and M. C. M. O'Brien, *The Jahn-Teller Effect in C_{60} and Other Icosahedral Complexes*, 1997; R. Englman, *The Jahn-Teller Effect in Molecules and Crystals*, 1972; T. A. Miller and V. E. Bondybey (eds.), *Molecular Ions*: *Spectroscopy*, *Structure and Chemistry*, 1983; Yu. Perlin and M. Wagner, *Modern Problems in Condensed Matter Sciences: The Dynamical Jahn-Teller Effect in Localized Systems*, 1985.

Jamming

Intentional generation of interfering signals in the electromagnetic and infrared spectrum by powerful transmitters as a countermeasure intended to block a communication, radar, or infrared system or to impair its effectiveness appreciably. Radio broadcasts or radio messages can be jammed by beaming a more powerful signal on the same frequency at the area in which reception is to be impaired, using carefully selected noise modulation to give maximum impairment of intelligibility of reception. When stations on many different frequencies are to be jammed or when an enemy is changing frequencies to avoid jamming, the jamming transmitter is correspondingly changed in frequency or swept through a range of frequencies over and over again. Techniques similar to this are also used at radar frequencies to jam early-warning and gunfire-control radar systems. *See* ELECTRONIC WARFARE. John Markus; Paul J. DeLia

Jasper

An opaque, impure type of massive fine-grained quartz that typically has a tile-red, dark-brownish-red, brown, or brownish-yellow color. The color of the reddish varieties of jasper is caused by admixed, finely divided hematite, and that of the brownish types by finely divided goethite. Jasper has been used since ancient times as an ornamental stone, chiefly of inlay work, and as a semiprecious gem material. Under the microscope, jasper generally has a fine, granular structure, but fairly large amounts of fibrous or spherulitic silica also may be present.

Jasper has a smooth conchoidal fracture with a dull luster. The specific gravity and hardness are variable, depending upon particle size and the nature and amount of the impurities present; both values approach those of quartz. The color of jasper often is variegated in banded, spotted, or orbicular types. Heliotrope is a translucent greenish chalcedony containing spots or streaks of opaque red jasper, and jaspagate contains bands of chalcedonic silica alternating with jasper. Jaspilite is a metamorphic rock composed of alternating layers of jasper with black or red bands of hematite. *See* CHALCEDONY; GEM; QUARTZ.

Clifford Frondel

Jaundice

The yellow staining of the skin and mucous membranes associated with the accumulation of bile pigments in the blood plasma. Bile pigments are the normal result of the metabolism of blood pigments, and are normally excreted from the blood into the bile by the liver. An increase in circulating bile pigments can, therefore, come about through increased breakdown of blood (hemolytic jaundice), through lack of patency of the bile ducts (obstructive jaundice), through inability or failure of the liver to clear the plasma (parenchymal jaundice), or through combinations of these. *See* GALLBLADDER; LIVER.

Metabolic pathway. Metabolism of the hemoglobin from destroyed red blood cells is carried on in organs of the reticuloendothelial system, such as the spleen, and the resulting bilirubin is liberated to the plasma. The plasma then circulates through the liver, where the bilirubin is conjugated enzymatically with glucuronic acid, and is excreted in the bile. Bile travels through the bile ducts to the small intestine, whence a small amount of altered bilirubin, termed urobilinogen, may be reabsorbed into the plasma. Excessive destruction of red blood cells causes accelerated production of bilirubin, overloads the ability of the liver to remove the pigment from the circulation, and produces jaundice. Blockage of the bile ducts causes elevation of plasma bilirubin glucuronide level because of the inability to dispose of this material in the usual channel. Damage to liver cells may cause elevation of the plasma bilirubin or bilirubin glucuronide or both, depending on the type and severity of liver cell damage.

Although the major portion of circulating bilirubin is derived from the breakdown of red cells, some is also contributed through inefficient or incomplete utilization of precursors of hemoglobin, which spill into the plasma bilirubin pool without having been used for hemoglobin synthesis. In addition, metabolism of cytochromes and similar pigmented compounds found in all cells of the body yields small quantities of bilirubin. *See* HEMOGLOBIN.

Jaundice occurs when the level of these circulating pigments becomes so high that they are visible in the skin and mucous membranes, where they are bound by a reaction which has not been identified. In the normal adult, levels of total bilirubin, that is, the total bilirubin and bilirubin glucuronide, rarely exceed 0.8–1.0 mg/100 ml of plasma, while jaundice usually becomes visible when total bilirubin approaches 1.5 mg. *See* BILIRUBIN.

Hemolytic jaundice. Destruction of red blood cells in the normal human adult proceeds at a rate at which about 0.8% of the circulating hemoglobin is broken down each day. This can be increased in states of excessive hemolysis up to 10- to 15-fold without overtaxing the remarkable ability of the liver to clear bilirubin from the plasma. Even this rate of clearing can be exceeded, however, in certain morbid states which result in hemolytic jaundice; such states include various hemolytic anemias, hemolysis resulting from incompatible blood transfusion, severe thermal or electric injuries, or introduction of hemolytic agents into the bloodstream. Similar jaundice occurs in pulmonary infarction.

In infants, and especially in premature infants, the ability of the liver to conjugate bilirubin with glucuronide is much less than in adults, apparently because of the lack of suitable enzymes. Jaundice appears in many infants shortly after birth, then disappears within a few days, with development of the appropriate enzyme structure. In the uncommon constitutional hepatic dysfunction, this enzyme defect apparently persists into adult life. The infantile jaundice accompanying erythroblastosis fetalis (Rh babies) is due to the inability of the infantile liver to metabolize the bilirubin resulting from markedly accelerated hemolysis. *See* BLOOD GROUPS; RH INCOMPATIBILITY.

A related form of jaundice occurs when an abnormality exists in the hemoglobin-forming cells of the bone marrow in which inadequate utilization of hemoglobin precursors occurs. Moderate degrees of jaundice can result from the accumulation of these substances in the bloodstream.

Obstructive jaundice. The highest levels of total bilirubin are seen in chronic obstructive jaundice, in which plasma levels may reach 50–60 mg/100 ml, and the skin may take on a remarkable deep-yellow hue. This condition may be brought about through a variety of means. In the infant there may be a severe maldevelopment of the bile ducts such that no channel for the flow of bile exists, while in the adult obstructive jaundice is most commonly caused by impaction of a gallstone in the ducts. Benign and malignant tumors of the gallbladder, bile ducts, pancreas, lymph nodes, and other organs may also cause compression of the bile ducts with loss of patency, and bile duct stricture may follow surgery or inflammation in the region. A similar, less severe, reversible picture is seen as a hypersensitivity response to the administration of some drugs, the most common of which are chlorpromazine and related drugs, and methyl testosterone. In the uncommon benign disorder known as idiopathic familial jaundice, there appears to be decreased ability to excrete conjugated bilirubin into the bile ducts, giving rise to a constant, but usually quite slight, elevation in the plasma bilirubin glucuronide. *See* GALLBLADDER DISORDERS.

Parenchymal jaundice. A wide variety of diseases exists in which part of the jaundice can be accounted for by actual damage to liver cells, with consequent decrease in their ability to conjugate bilirubin and excrete the glucuronide, causing an elevation of both fractions in the plasma. This group comprises such conditions as inflammations of the liver, including viral hepatitis, Weil's disease, syphilis, parasitic infestations, and bacterial infections; toxic conditions, including poisoning from a wide variety of organic and inorganic compounds and in a broader sense the toxemias associated with severe systemic diseases; tumorous conditions, including primary hepatic tumors and those metastatic from other organs; and other conditions, the most common of which is congestive heart failure. Some of these conditions have an added component of obstructive or hemolytic jaundice, which confuses the picture for the clinician.

Symptomatology. The appearance and symptomatology of subjects suffering from jaundice vary from case to case and depend on the underlying disease process producing the jaundice. Elevation of bilirubin by itself has very limited deleterious effects on overall physiology, with two exceptions: The brain of very young infants is subject to damage by high levels of circulating bilirubin (a condition termed kernicterus); and, under certain conditions in the adult, high levels of circulating bilirubin appear to contribute to kidney damage.

Aside from the effects of the jaundice itself, however, individuals with hemolytic jaundice usually have an accompanying anemia. Those with obstructive jaundice commonly note that their stools are not brown, a symptom caused by lack of bile pigments (acholic stools), while bilirubin glucuronide appears in the urine and causes it to turn dark. Persons with malignancies demonstrate the usual signs of weight loss and anemia, while those with inflammatory conditions commonly have fever, chills, and prostration. *See* ANEMIA; PIGMENTATION.

Rolla B. Hill

Jawless vertebrates

The common name for members of the agnathans. Jawless vertebrates include the cyclostomes (modern lampreys and hagfishes) as well as extinct armored fishes, known colloquially as ostracoderms ("bony-skinned"), that lived in the Ordovician, Silurian, and Devonian periods (**Fig. 1**). Agnathans have pouchlike gills opening through small pores, rather than slits as in jawed vertebrates (Gnathostomata). Primitively, agnathans lack jaws, they show a persistent notochord, and most have no paired fins. *See* CYCLOSTOMATA (CHORDATA).

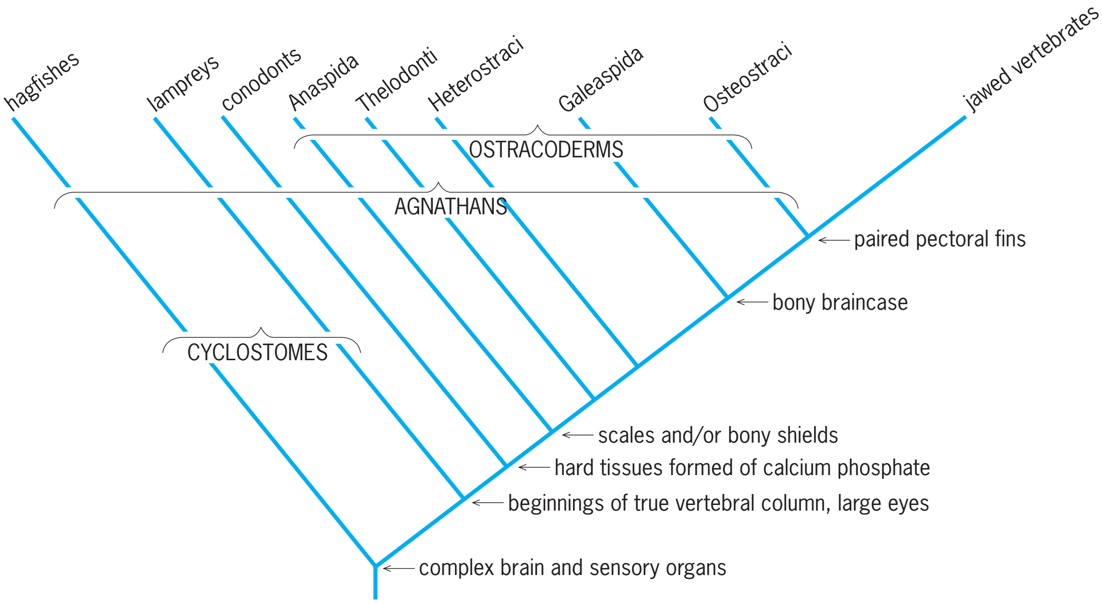
Types. The lampreys (Petromyzontiformes; 38 extant species) and hagfishes (Myxiniformes; $\sim$70 species) are scaleless, eel-shaped fishes with round mouths, inside which are keratinized teeth carried upon a complex tongue and oral hood. Parasitic lampreys attach to host fishes by a sucker and use the tongue to rasp away flesh and blood. Hagfishes are exclusively marine, are blind, and live most of their lives buried in mud, emerging to eat polychaete worms, mammal carcasses, or dead or dying fishes. Lampreys and hagfishes show a bipolar distribution (they occur in both Northern and Southern hemispheres, but only 4 species occur in the Southern) and prefer cool waters. *See* MYXINIFORMES; PETROMYZONTIFORMES.

Ostracoderms were very variable in shape and were covered with a superficial bony armor made up of solid shields or scales. One of the best known was the osteostracans (cephalaspids) of Eurasia and North America, with a solid semicircular head shield pierced by dorsally placed eyes and a small circular mouth on the undersurface. The head shield is also marked by sensory fields (lacunae filled with small plates) which were specialized parts of the lateral-line system. Osteostracans probably lived on, or were partly buried in, mud and sand, where they sucked in small food particles. *See* CEPHALASPIDOMORPHA; OSTEOSTRACI; OSTRACODERM.

Galeaspids were superficially like osteostracans and probably occupied a similar ecological niche in China. The head shield in many was extended into an elaborate snout.

Heterostracans (pteraspids) were torpedo-shaped active swimmers that lived alongside osteostracans. The head was encased in large dorsal and ventral shields which surrounded small, laterally directed eyes. The mouth, formed by a lower lip equipped with small bony plates, may have been extended as a scoop to ingest food particles. A few ostracoderms superficially look like heterostracans and sometimes are classified with them because they show a similar type of bone which lacks enclosed osteocytes (bone cells). Heterostracans are the oldest of completely preserved vertebrates and lived in seas 460 million years ago. *See* HETEROSTRACI.

The Thelodonti are a group of small ostracoderms in which the body is completely covered with



Interrelationships of jawless and jawed vertebrates, showing distribution of some major evolutionary steps.

minute, finely sculptured scales. These scales are found worldwide and in great abundance in Upper Silurian and Lower Devonian rocks. Consequently, they have been used for stratigraphic zonation and correlation. Some scientists think that thelodonts may be closely related to sharks and rays (Chondrichthyes). *See* CHONDRICHTHYES.

Interrelationships. Modern ideas of the interrelationships of agnathan fishes suggest that they are not a natural group; that is, some such as the osteostracans and galeaspids are more closely related to jawed vertebrates than to other agnathans. Similarly, lampreys share many specializations with jawed vertebrates not seen in hagfishes, such as neural arches along the notochord, eye muscles, nervous control of heartbeat, and the capability to osmoregulate and adapt to freshwater. These attributes suggest that lampreys are more closely related to jawed vertebrates than to hagfishes and that the Cyclostomata are not a natural group.

Recently, Conodonta have been classified as vertebrates. Conodonts are preserved as tiny, phosphatic, toothlike elements found in marine deposits from the Cambrian to Triassic. Usually found in great abundance, they are useful for dating and correlating rocks. Rare fossils of soft-bodied, eel-shaped animals show that conodonts are part of complex feeding devices within the mouth. They grew to at least 5 cm (2 in.) (some incompletely known forms must have been considerably longer) and had large eyes, suggesting an active predatory lifestyle. *See* CONODONT.

A classification of jawless vertebrates is given below:

- Subphylum Craniata (complex brain and sensory organs)
  - Class Hyperotreti
    - Order Myxiniformes
  - Class Vertebrata (with neural arches in backbone)
    - Order: Petromyzontiformes
      - Conodonta
      - Anaspida
      - Thelodonti
      - Heterostraci (pteraspids)
      - Galeaspida
      - Osteostraci (cephalaspids)

Peter L. Forey

Bibliography. P. C. J. Donoghue et al., Conodont affinity and chordate phylogeny, *Biol. Rev.*, 75:191–251, 2000; P. L. Forey and P. Janvier, Evolution of the early vertebrates, *Amer. Sci.*, pp. 554–565, November–December 1994; P. Janvier, *Early Vertebrates*, Clarendon Press, Oxford, 1996.

Jerboa

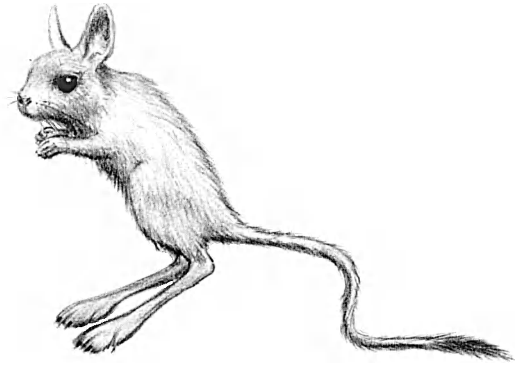
The common name for 33 species of rodents in the family Dipodidae (see **table**). Jerboas inhabit the hot arid desert and semiarid regions of Turkey, North Africa, and Asia. *See* DESERT; RODENTIA.

TABLE 1. Genera and common names of jerboas

Genus	Common name
Cardiocranius	Five-toed dwarf jerboa (1 species)
Salpingotus	Three-toed dwarf jerboas (5 species)
Salpingotulus	Baluchistan pygmy jerboa (1 species)
Paradipus	Comb-toed jerboa (1 species)
Euchoreutes	Long-eared jerboa (1 species)
Dipus	Rough-legged jerboa, or northern three-toed jerboa (1 species)
Eremodipus	Central Asian jerboa (1 species)
Jaculus	Desert jerboas (4 species)
Stylodipus	Thick-tailed three-toed jerboas (3 species)
Allactaga	Four- and five-toed jerboas (11 species)
Allactodipus	Bobrinski's jerboa (1 species)
Alactagulus	Lesser five-toed jerboa, or little earth hare (1 species)
Pygeretmus	Fat-tailed jerboas (2 species)

Description. Jerboas are rodents ranging 2–8 in. (50–200 mm) in head and body length (see **illustration**). The rounded tail is usually about 2 in. (50 mm) longer than the head and body. The ears are large, equaling half the length of the head and body in some species. The eyes are large. Coloration varies but is usually some shade of buff (moderate orange-yellow) mixed with black or russet; the underparts are white.

Of all the rodents, this kangaroolike mammal is perhaps the most highly developed for getting around on two feet. The hindlimbs are large and very long (at least four times the length of the front limbs). In most species the three main foot bones are fused into a single "cannon bone" for greater strength. The front limbs are considerably smaller and are used primarily for grooming and for holding the grass seeds on which many species feed. Jerboas prefer to stand up on their hindlegs and hop, covering 2–6 ft (0.6–1.8 m) in a single jump. When alarmed, a jerboa can travel by leaps and bounds of 5–10 ft (1.5–3 m), considerably faster than a human can run. *Allactaga elater* has been timed at a speed of 48 km/h (30 mi/h). The tail, which ends in a tuft of hair, not only gives the jerboa support when it is standing but also helps the animal maintain its balance as it leaps along. The thick hairs on the feet absorb some of the landing shock and provide traction on the sand.



Typical jerboa.

The largest of all jerboas is the four-toed jerboa (*Allactaga major*). Head and body length may range 90–263 mm (3.5–10 in.), tail length is 142–308 mm (5.5–12 in.), and weight is 280–420 grams (9.75–14.75 oz).

Behavior. Jerboas are nocturnal in order to escape the intense heat of the day. They spend the day holed up in their underground den. When they retire for the day, they close up the burrow entrance to keep out the hot sunlight. Four types of burrows are utilized by different species of jerboas: temporary summer day burrows, temporary summer night burrows, permanent summer burrows used as living quarters and for producing young, and permanent winter burrows for hibernation. Temporary shelters during the day may consist of a long but simple tunnel up to 13 m (42 ft) in length and 20–30 cm (8–12 in.) deep. Some species are solitary, while others are more social.

Jerboas in Russia, Ukraine, and Central Asia may hibernate for 6–7 months, whereas jerboas in more southern regions may not hibernate at all. Some species feed primarily on seeds and other plant material such as bulbs, roots, and rhizomes (underground horizontal stems); others are mainly insectivorous. Some fluids are obtained through their food, but jerboas also produce fluid internally as a result of metabolic processes (“metabolic water”). *See* METABOLISM.

Breeding extends from about March to October with females giving birth to two or three litters annually. Most litters contain three to five young. The gestation period is 25–40 days. Females of some species attain sexual maturity in $3^1/_2$ months; therefore, a female born early in the breeding season could give birth to her own litter during the same breeding season.

Several species of jerboas are classified as endangered, threatened, or vulnerable by the International Union for the Conservation of Nature and Natural Resources (IUCN) because of their restricted range and/or habitat disturbance. Donald W. Linzey

Bibliography. D. Macdonald (ed.), *The Encyclopedia of Mammals*, Andromeda Oxford, 2001; R. M. Nowak, *Walker’s Mammals of the World*, 6th ed., Johns Hopkins University Press, 1999.

Jet (gemology)

A black, opaque material that takes a high polish. Jet has been used for many centuries for ornamental purposes. It is a compact variety of lignite coal. It has a refractive index of 1.66 (which accounts for its relatively high luster), a hardness of 3–4 on Mohs scale, and a specific gravity of 1.30–1.35. Jet is compact and durable, and can be carved or even turned on a lathe. The principal source is Whitby, England, where it occurs in hard shales. Although popular from Pliny’s day until the nineteenth century, the use of jet for jewelry purposes declined markedly for centuries until a resurgence of interest occurred in the 1960s. *See* GEM; LIGNITE. Richard T. Liddicoat, Jr.

Jet flow

A fluid flow in which a stream of one fluid mixes with a surrounding medium, at rest or in motion. Such flows occur in a wide variety of situations, and the geometries, sizes, and flow conditions cover a large range. To create some order, two dimensionless groupings are used. The first is the Reynolds number, defined in Eq. (1), where ρ is the density, V

$$\mathrm{Re} \equiv \frac{\rho V L}{\mu} \qquad (1)$$

is a characteristic velocity (for example, the jet exit velocity), L is a characteristic length (for example, the jet diameter), and μ is the viscosity. The second is the Mach number, defined in Eq. (2), where a is

$$\mathrm{M} \equiv \frac{V}{a} \qquad (2)$$

the speed of sound. Jet flows vary greatly, depending on the values of these two numbers. *See* MACH NUMBER; REYNOLDS NUMBER.

Laminar and turbulent flows. Jets exiting into a similar fluid (liquid into liquid, or gas into gas) almost always become turbulent within a short distance of their origin, irrespective of whether the flow in the supply line is laminar or turbulent. For example, a water jet exiting into water at rest with $\mathrm{Re} \approx 2300$ (**Fig. 1**) is initially in the simple laminar state, but at

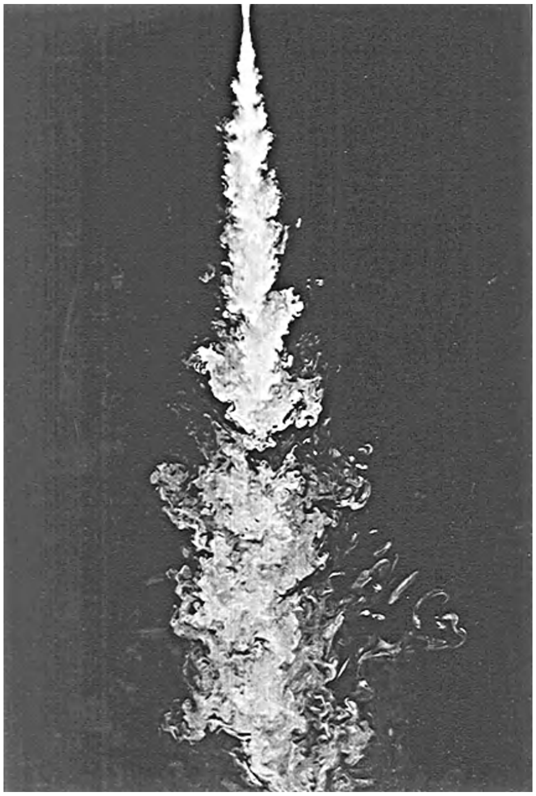


Fig. 1. Laser-induced fluorescence of a round dyed water jet directed downward into water at rest, showing the instantaneous concentration of injected fluid in the plane of the jet axis. The Reynolds number is approximately 2300. (*From P. E. Dimotakis, R. C. Lye, and D. Z. Papantoniou, Proceedings of the 15th International Symposium on Fluid Dynamics, Jachranka, Poland, 1981*)

this Reynolds number that state is unstable and the flow undergoes a transition to the more chaotic turbulent state. Turbulent structures called eddies are formed with a large range of sizes. The large-scale structures are responsible for capturing fluid from the surroundings and entraining it into the jet. However, the jet and external fluids are not thoroughly mixed until diffusion is completed by the small-scale structures. A more stable case, which may remain laminar for a long distance, is where a liquid jet issues into air or other light gas. *See* DIFFUSION; LAMINAR FLOW; TURBULENT FLOW.

High-speed flows. When the velocities in the jet are greater than the speed of sound ($M > 1$) the flow is said to be supersonic, and important qualitative changes in the flow occur. The most prominent change is the occurrence of shock waves. For example, a supersonic air jet exhausting from a nozzle at low pressure into higher-pressure air at rest is said to be overexpanded. Such a flow can be visualized by an optical technique called schlieren photography (**Fig. 2**). As the jet leaves the nozzle, it senses the higher pressure around it and adjusts through oblique shock waves emanating from the edges of the nozzle. As the waves approach the jet axis, they are focused and strengthen to form a vertical Mach disk, visible in the middle of the jet. The pattern repeats itself until dissipated by viscous effects. If the supersonic jet exits the nozzle at higher pressure than the surroundings, it will plume out. Such behavior is observed as the space shuttle rises to high altitudes with the main engines running. *See* SCHLIEREN PHOTOGRAPHY; SHOCK WAVE; SUPERSONIC FLOW.

Plumes. Another class of jet flows is identified by the fact that the motion of the jet is induced primarily by buoyancy forces. A common example is a hot gas exhaust rising in the atmosphere. Such jet flows are called buoyant plumes, or simply plumes, as distinct from the momentum jets, or simply jets, discussed above. These flows become very interesting when the surrounding fluid is stratified in different density layers, as can happen in the atmosphere and ocean by heating from the Sun. A heated plume may then rise rapidly until it reaches a layer of gas at the same density, where the buoyancy force due to temperature difference will drop to zero. The plume will then spread out in that layer without further rise. Such situations can be important for pollution-dispersal problems. *See* AIR POLLUTION; BOUNDARY-LAYER FLOW; FLUID FLOW.

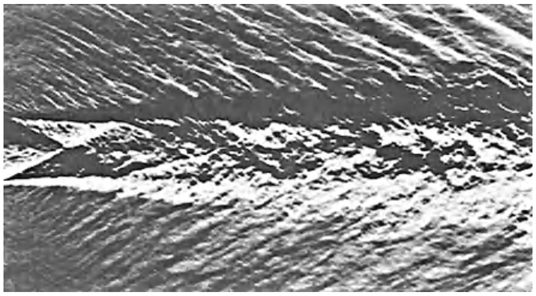


Fig. 2. Schlieren photograph of an overexpanded Mach 1.8 round air jet exiting from left to right into ambient air. The adjustment of the lower pressure in the surroundings leads to oblique shock waves and Mach disks in the jet. (*From H. Oertel, Modern Developments in Shock Tube Research, Shock Tube Research Society, Japan, 1975*)

Joseph A. Schetz

Bibliography. G. N. Abramovich, *Theory of Turbulent Jets*, 1963; J. A. Schetz, *Boundary Layer Analysis*, 1993; J. A. Schetz, *Injection and Mixing in Turbulent Flow*, 1980.

Jet stream

A relatively narrow, fast-moving wind current flanked by more slowly moving currents. Jet streams are observed principally in the zone of prevailing westerlies above the lower troposphere and in most cases reach maximum intensity, with regard both to speed and to concentration, near the tropopause. At a given time, the position and intensity of the jet stream may significantly influence aircraft operations because of the great speed of the wind at the jet core and the rapid spatial variation of wind speed in its vicinity. Lying in the zone of maximum temperature contrast between cold air masses to the north and warm air masses to the south, the position of the jet stream on a given day usually coincides in part with the regions of greatest storminess in the lower troposphere, though portions of the jet stream occur over regions which are entirely devoid of cloud. The jet stream is often called the polar jet, because of the importance of cold, polar air. The subtropical jet is not associated with surface temperature contrasts, like the polar jet. Maxima in wind speed within the jet stream are called jet streaks. *See* CLEAR-AIR TURBULENCE.

Characteristics. The specific characteristics of the jet stream depend upon whether the reference is to a single instantaneous flow pattern or to an averaged circulation pattern, such as one averaged with respect to time, or averaged with respect both to time and to longitude.

If the winter circulation pattern on the Northern Hemisphere is averaged with respect to both time and longitude, a westerly jet stream is found at an elevation of about 8 mi (13 km) near latitude (lat) 25°. The speed of the averaged wind at the jet core is about 80 knots (148 km/h). In summer the jet is displaced poleward to a position near lat 41°. It is found at an elevation of about 7 mi (12 km) with a maximum speed of about 30 knots (56 km/h). In both summer and winter a speed equal to one-half the peak value is found approximately 15° of latitude south, 20° of latitude north, 3–6 mi (5–10 km) above, and 3–6 mi (5–10 km) below the location of the jet core itself.

If the winter circulation is averaged only with respect to time, it is found that both the intensity and the latitude of the westerly jet stream vary from one sector of the Northern Hemisphere to another. The most intense portion, with a maximum speed of about 100 knots (185 km/h), lies over the extreme western portion of the North Pacific Ocean at about lat 22°. Lesser maxima of about 85 knots (157 km/h) are found at lat 35° over the east coast of North America, and at lat 21° over the eastern Sahara and over the Arabian Sea. In summer, maxima are found at lat 46° over the Great Lakes region, at lat 40° over the western Mediterranean Sea, and at lat 35° over the central North Pacific Ocean. Peak speeds in these regions range between 40 and 45 knots (74 and 83 km/h). The degree of concentration of these jet streams, as measured by the distance from the core to the position at which the speed is one-half the core speed, is only slightly greater than the degree of concentration of the jet stream averaged with respect to time and longitude. At both seasons at all longitudes the elevation of these jet streams varies between 6.5 and 8.5 mi (11 and 14 km).

Variations. On individual days there is a considerable latitudinal variability of the jet stream, particularly in the western North American and western European sectors. It is principally for this reason that the time-averaged jet stream is not well defined in these regions. There is also a great day-to-day variability in the intensity of the jet stream throughout the hemisphere. On a given winter day, speeds in the jet core may exceed 200 knots (370 km/h) for a distance of several hundred miles along the direction of the wind. Lateral wind shears in the direction normal to the jet stream frequently attain values as high as 100 knots per 300 nautical miles (185 km/h per 556 km) to the right of the direction of the jet stream current and as high as 100 knots per 100 nautical miles (185 km/h per 185 km) to the left. Vertical shears below and above the jet core as often as large as 20 knots per 1000 ft (37 km/h per 305 m). Daily jet streams are predominantly westerly, but northerly, southerly, and even easterly jet streams may occur in middle or high latitudes when ridges and troughs in the normal westerly current are particularly pronounced or when unusually intense cyclones and anticyclones occur at upper levels.

Insufficiency of data on the Southern Hemisphere precludes a detailed description of the jet stream, but it appears that the major characteristics resemble quite closely those of the jet stream on the Northern Hemisphere. The day-to-day variability of the jet stream, however, appears to be less on the Southern Hemisphere.

It appears that an intense jet stream occurs at high latitudes on both hemispheres in the winter stratosphere at elevations above 12 mi (20 km). The data available, however, are insufficient to permit the precise location or detailed description of this phenomenon. *See* AIR MASS; ATMOSPHERE; ATMOSPHERIC WAVES, UPPER SYNOPTIC; GEOSTROPHIC WIND; STORM; VORTEX

. Frederick Sanders; Howard B. Bluestein

Jewel bearing

A bearing used in quality timekeeping devices, gyros, and instruments; usually made of synthetic corundum (crystallized Al_2O_3), which is more commonly known as ruby or sapphire. The extensive use of such bearings in the design of precision devices is mainly due to the outstanding qualities of the material. Sapphire's extreme hardness imparts to the bearing excellent wear resistance, as well as the ability to withstand heavy loads without deformation of shape or structure. The crystalline nature of sapphire lends itself to very fine polishing and this, combined with the excellent oil- and lubricant-retention ability of the surface, adds to the natural low-friction characteristics of the material. Sapphire is also nonmagnetic and oxidization-resistant, and has a very high melting point (3685°F or 2029°C). Ruby has the same properties as sapphire; the red coloration is due to the introduction of a small amount of chromium oxide. *See* ANTIFRICTION BEARING; GEM; GYROSCOPE; RUBY; SAPPHIRE; WATCH.

Types. Jewel bearings, classified as either instrument or watch jewels, are also categorized according to their configuration or function. The ring jewel is the most common type. It is basically a journal bearing which supports a cylindrical pivot. The wall of the hole can be either left straight (bar hole) or can be imparted a slight curvature from end to end (olive hole). This last configuration is designed to reduce friction, compensate for slight misalignment, and help lubrication. A large variety of instrument and timing devices are fitted with such bearings, including missile and aircraft guidance systems. *See* GUIDANCE SYSTEMS.

Vee, or V, jewels are used in conjunction with a conical pivot, the bearing surface being a small radius located at the apex of a conical recess. This type of bearing is found primarily in electric measuring instruments.

Cup jewels have a highly polished concave recess mated to a rounded pivot or a steel ball. Typical are compass and electric-meter bearings.

End stone and cap jewels, combined with ring jewels, control the end play of the pivot and support axial thrust. They consist of a disk with highly polished flat or convex ends. Other relatively common jewel bearings are pallet stones and roller pins; both are part of the time- keeping device's escapement.

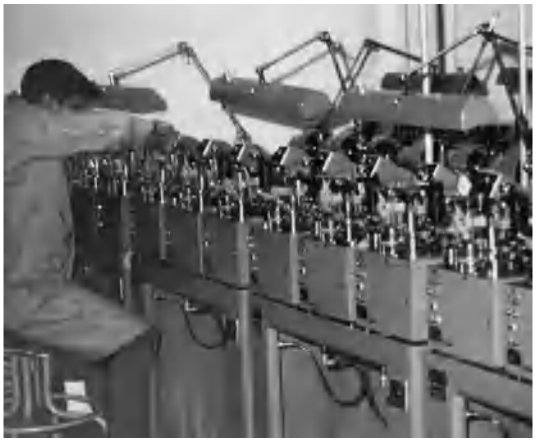


Fig. 1. Automatic cupping machines for the manufacture of jewel bearings. (*Bulova Watch Co.*)

Dimensions. Minute dimensions are a characteristic of jewel bearings. A typical watch jewel may be 0.040 in. (0.10 cm) in diameter with a 0.004-in. (0.01-cm) hole, but these dimensions may go down to 0.015 and 0.002 in. (0.038 and 0.005 cm), respectively. Jewels with a diameter of more than $^1/_{16}$ in. (0.16 cm) are considered large. It is usual for critical dimensions, such as hole diameter and roundness, to have a tolerance of 0.0001 in. (0.00025 cm) or less. In some instances these tolerances may be as low as 0.0000020 in. (0.000005 cm).

Manufacturing. Because of its hardness sapphire can only be worked by diamond, which is consequently the main tool for the production of jewel bearings. Both natural and synthetic diamond are used, mostly under the form of slurry, broaches, and grinding wheels. *See* DIAMOND.

The machining of the blanks, small disks, or cylinders of varied diameter and thickness is the first step in the manufacturing process, and is common to most types of jewel bearings. The boules (pear-shaped crystals of synthetic ruby or sapphire) are first oriented according to the optical axis to ensure maximum hardness of the bearing working surface. They are then sliced, diced, and ground flat, and rounded by centerless grinding to the required blank dimensions. From this point on, the process varies considerably, according to the type of bearing.

Ring jewels are drilled with a steel or tungsten wire and coarse diamond slurry, or bored with a small grinding tool. The hole is then enlarged and sized by reciprocating wires of increasingly larger sizes through a string of jewels, until the required hole size is achieved. Fine diamond powder and a very slow rate of material removal permit the respect of strict tolerances and high-finish quality requirements. The jewels, supported by a wire strung through the hole, are then ground in a special centerless-type grinding machine to the desired outside diameter dimension. After the cutting of a concave recess, which functions as an oil reservoir, the top and bottom of the bearing are polished and beveled by lapping and brushing. Finally, the "olive" configuration is obtained by oscillating a diamond charged wire through the hole of the rotating jewel. Between each operation the boiling of the jewels in a bath of sulfuric and nitric acid disposes of remaining slurries and other contaminating products.

Fig. 2. Individual head of an automatic cupping machine. (*Bulova Watch Co.*)

Fig. 3. Technician shown operating an automatic cupping machine. (*Bulova Watch Co.*)

The conical recess in vee jewels is first roughly shaped by a pyramidal diamond tool. The wall of the vee and the radius are then polished and blended with an agglomerated diamond broach, and a succession of brushing operations. Lapping of the top of the jewel and brushing a bevel around the upper outside edge conclude the process.

Most other types of jewel bearings, such as end stones and pallet stones, are shaped by a series of grinding, lapping, and brushing operations.

A full line of automatic and semiautomatic high-precision equipment has been developed to handle and machine jewel bearings efficiently, permitting mass production and relatively low cost (**Figs. 1–3**). Traditionally, a large proportion of the labor involved is devoted to in-process and final quality control.

Robert M. Schultz

Johne's disease

A slowly progressive diarrheal disease that causes major economic loss to the cattle industry; also known as paratuberculosis. It is caused by *Mycobacterium avium ss paratuberculosis (MAP)*, which produces chronic inflammation of the mucosa of the ileocecal valve and adjacent tissues of the gastrointestinal tract of cattle, sheep, goats, and wild ruminants. The organism has been isolated from 2–3% of the adult cows in the United States; in some exotic species (such as mouflon and mountain goats), more than 90% of the animals in a herd may be infected. Animals less than 2 months of age are most susceptible to infection. Therefore, eliminating or minimizing exposure of young animals to *MAP* is important in the control of Johne's disease. It is recommended that calves be removed from dams immediately after birth and fed colostrum from negative animals. *See* MYCOBACTERIAL DISEASES.

Transmission. Transmission of *MAP* is primarily by ingestion of feces from animals shedding the organism. The incubation period varies from 1 to 3 years or more. Diseased animals in early or subclinical stages have intermittent or persistent diarrhea without fever and in advanced stages become emaciated. Congenital infections have been reported, and the organism has been isolated from colostrum, supramammary lymph nodes, fetal tissues, blood, and semen of cattle with clinical disease.

Diagnosis. Johne's disease is diagnosed by mycobacteriologic examinations conducted on feces or tissues collected by biopsy or at necropsy. Enzyme-linked immunosorbent assays (ELISA) have been developed using a purified protein derivative prepared from the culture filtrate of *MAP* or chemical extracts of the bacterium, including proteins and carbohydrates (antigens) for detecting antibodies in sera of animals exposed to the organism. ELISA reportedly fails to identify three of four contagious cows shedding *MAP* in feces. Moreover, animals in the early stages of disease are frequently negative for humoral antibodies on ELISA, complement fixation, and/or agar gel immunodiffusion tests. Therefore, to confirm a diagnosis of Johne's disease, it is necessary to isolate and identify the etiologic agent by mycobactin dependency and/or polymerase chain reaction (PCR).

Vaccination. A killed whole-cell vaccine available for use in calves 1–35 days of age markedly reduces and often eliminates the occurrence of clinical disease in adult cattle. Live attenuated strains of *MAP* have been used for vaccinating cattle in a few countries, but are not approved for use in the United States or Canada. Therapeutic drugs are not available for routine treatment of animals.

Relation to Crohn's disease. The superficial similarity of Johne's disease to Crohn's disease in humans has led to the hypothesis that Crohn's disease is also caused by *MAP*. Although *MAP* has been isolated from some Crohn's disease patients, no definitive information on association is available. Moreover, published information on Crohn's disease in humans fails to show any association with bovine paratuberculosis.

Available information indicates that the parasitic worm *Trichuris suis* may be useful in the treatment of patients with Crohn's disease. This helminth may play a role in altering host mucosal immunity by inhibiting dysregulated inflammatory responses. *See* INFLAMMATORY BOWEL DISEASE.

Charles O. Thoen

Bibliography. O. Chacon, L. E. Bermudez, and R. G. Baretta, Johne's disease, inflammatory bowel disease and *Mycobacterium paratuberculosis*, *Annu. Rev. Microbiol.*, 58:329–363, 2004; P. H. Jones et al., Crohn's disease in people exposed to clinical cases of bovine paratuberculosis, *Epidemiol. Infect.*, pp. 1–8, January 2005; R. W. Summers et al., *Trichuris suis* therapy in Crohn's disease, *Gut*, 54:87–89, 2005; C. O. Thoen and R. G. Barletta, *Mycobacterium*, in *Pathogenesis of Bacterial Infections in Animals*, pp. 70–77, Blackwell, 2004; C. O. Thoen and J. A. Haagsma, Molecular techniques in the diagnosis and control of paratuberculosis in cattle, *J. Amer. Vet. Med. Ass.*, 209:734–737, 1996; J. V. Weinstock, R. W. Summers, and D. E. Elliott, Role of helminths in regulating mucosal inflammation, *Springer Seminars Immunopathol.*, 27:249–271, 2005.

Joint (anatomy)

The structural component of an animal skeleton where two or more skeletal elements meet, including the supporting structures within and surrounding it. The relative range of motion between the skeletal elements of a joint depends on the type of material between these elements, the shapes of the contacting surfaces, and the configuration of the supporting structures.

In bony skeletal systems, there are three general classes of joints: synarthroses, amphiarthroses, and diarthroses. Synarthroses are joints where bony surfaces are directly connected with fibrous tissue, allowing very little if any motion. Synarthroses may be further classified as sutures, syndesmoses, and gomphoses. Sutures are joined with fibrous tissue, as in the coronal suture where the parietal and frontal bones of the human skull meet; in adulthood, this suture becomes an essentially rigid bony union called a synostosis. Syndesmoses are connected with ligaments, as are the shafts of the tibia and fibula. The roots of a tooth that are anchored in the jaw bone with fibrous tissue form a gomphosis. Amphiarthroses are joints where bones are directly connected with fibrocartilage or hyaline cartilage and allow only limited motion. An amphiarthrosis joined with fibrocartilage, as found between the two pubic bones of the pelvis, is known as a symphysis; but when hyaline cartilage joins the bones, a synchondrosis is formed, an example being the first sternocostal joint. The greatest range of motion is found in diarthrodial joints, where the articulating surfaces slide and to varying degrees roll against each other. *See* LIGAMENT.

Diarthrodial joints. The contacting surfaces of the bones of a diarthrodial joint are covered with articular cartilage, an avascular, highly durable hydrated soft tissue that provides shock absorption and lubrication functions to the joint (**Fig. 1**). Articular cartilage is composed mainly of water, proteoglycans, and collagen. The joint is surrounded by a fibrous joint capsule lined with synovium, which produces lubricating synovial fluid and nutrients required by the tissues within the joint. Joint motion is provided by the muscles that are attached to the bone with tendons. Strong flexible ligaments connected across the bones stabilize the joint and may constrain its motion.

Different ranges of motion result from several basic types of diarthrodial joints: pivot, gliding, hinge, saddle, condyloid, and ball-and-socket. *See* COLLAGEN.

Pivot. Pivot (trochoid) joints allow rotation of one bone about another, as does the median atlantoaxial joint (atlas about the second cervical vertebra) at the base of the skull.

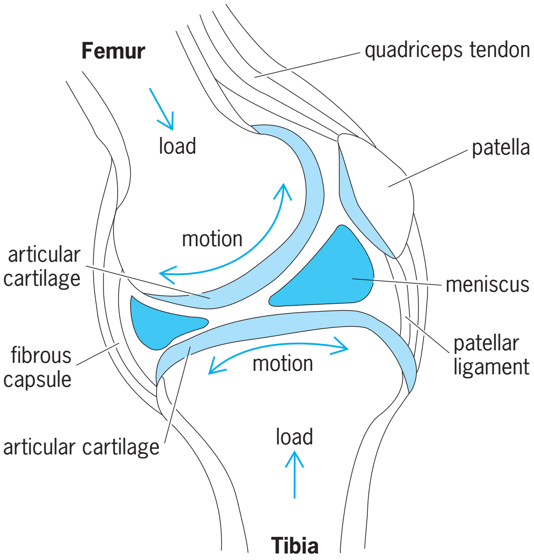


Fig. 1. Cross section of the human knee showing its major components. This diarthrodial joint contains contacting surfaces on the tibia, femur, meniscus, and patella (knee cap). The patella protects the joint and also serves to redirect the force exerted by the quadriceps muscles to the tibia. (*After R. Skalak and S. Chien, eds., Handbook of Bioengineering, McGraw-Hill, 1987*)

Gliding. Some of the joints found in the wrist and ankle are examples of gliding (arthrodia) joints. The surfaces of these joints slide against one another and provide limited motion in multiple directions.

Hinge. Although the hinge (ginglymus) joint generally provides motion in only one plane (flexion and extension), its motion, as in the human knee, can be complex. The articulating surfaces of the knee both roll and slide past each other, and the motion of the tibia relative to the femur displays a small amount of rotation. The knee also contains two fibrocartilaginous (primarily collagenous) crescent-shaped discs between the surfaces of the tibia and femur, called menisci, which provide shock absorption, stability, and lubrication functions to the knee (Fig. 1). The elbow is another example of a hinge joint.

Saddle. This joint is found at the base of the human thumb between the carpal and metacarpal (trapezium). It is characterized by saddle-shaped (concave and convex) articulating surfaces. The saddle joint provides a large range of motion (flexion, extension, abduction and adduction) with limited rotation.

Condyloid. The range of motion of the condyloid joint is similar to that of the saddle joint. Here, the articulating surfaces are ovoid, and one surface is convex, the other concave. This joint is found at the base of the fingers where the metacarpal and proximal phalanx meet. The surfaces of the knee joint are also of condyloid shapes (**Fig. 2**).

Ball-and-socket. The human shoulder and hip are examples of the ball-and-socket (enarthrosis) joint. In addition to flexion, extension, abduction, and adduction, they provide rotation about the long axis of the bone. The low relative-surface-area ratio of the glenoid (socket) to humeral head (ball) of the shoulder gives it the largest range of motion of all human joints. This high degree of mobility is stabilized by a complex rotator cuff consisting of ligaments and muscular attachments. The larger ratio of acetabulum (socket) to femoral head (ball) surface area of the hip constrains the joint to lesser mobility but provides it with greater inherent stability.

Lubrication. The lubrication mechanisms of diarthrodial joints are extremely effective. Although the surface roughness of articular cartilage is higher than that of some engineered bearing surfaces, the coefficient of friction between cartilage surfaces is significantly lower than between most manufactured materials. The rate of wear of the articular surfaces is also extremely low in the normally functioning joint. While many details of diarthrodial joint lubrication mechanisms are not fully understood, these mechanisms fall into two categories: fluid-film lubrication, where the joint surfaces are separated by a layer of lubricating fluid (synovial fluid); and boundary lubrication, where the joint surfaces are in contact with each other. These mechanisms probably occur together to varying extents, depending upon the particular joint and the type of motion.

Some of the major theories involving fluid-film lubrication are hydrodynamic, squeeze film, hydrostatic, and boosted lubrication.

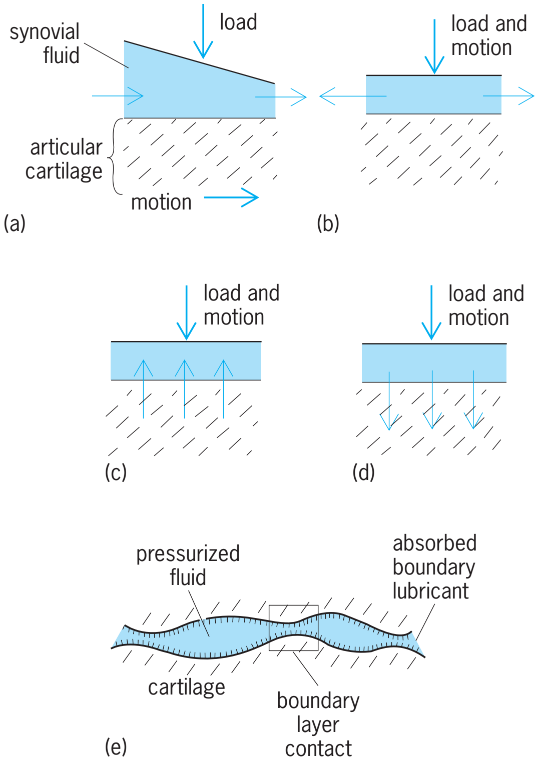


Fig. 2. Possible modes of lubrication in diarthrodial joints. (*a–d*) Different types of fluid-film lubrication showing the directions of loading, motion, and fluid flow: (*a*) hydrodynamic; (*b*) squeeze film; (*c*) hydrostatic weeping; (*d*) boosted. (*e*) A combination of boundary layer lubrication at points of contact and fluid-film lubrication. (*After V. C. Mow and W. C. Hayes, eds., Basic Orthopaedic Biomechanics, Raven Press, 1991*)

Hydrodynamic. If the relative velocity between the joint surfaces is sufficiently high, a pressurized, wedge-shaped layer of the intervening synovial fluid may be formed (Fig. 2*a*). This fluid layer generates lift which forces the joint surfaces apart. Hydrodynamic lubrication probably does not occur to any great extent in diarthrodial joints, except under conditions of low load and high velocity that might occur, for example, during high-speed pitching of a ball.

Squeeze film. When both surfaces of a joint approach each other, pressure in the synovial fluid is generated because this viscous lubricant cannot instantaneously be squeezed from the gap (Fig. 2*b*). The resulting pressurized squeeze film can support large loads which may be generated, for example, in the knee joint when jumping from a height. Since the articular cartilage of the joint surfaces is deformable, the synovial fluid may become trapped in localized depressions. This film may support large loads for many minutes before becoming depleted.

Hydrostatic. During normal joint loading, compression of the articular cartilage surfaces may result in the weeping of fluid from the cartilage into the joint cavity (Fig. 2*c*). Weeping lubrication, a form of hydrostatic lubrication, is facilitated by this self-pressurization of the fluid within the cartilage.

Boosted. As the surfaces of the joint approach each other during normal function, it is possible that the resulting pressure causes some of the water in the synovial fluid to be forced into the cartilage (Fig. 2*d*). Boosted lubrication results from the remaining, highly viscous synovial fluid concentrate between the two articulating surfaces.

Boundary. Boundary layer lubrication takes place when the surfaces of the joints are touching each other. It has been proposed that a monolayer of a polypeptide chain, called lubricin, is adsorbed onto the cartilage surface. This boundary lubricant separates the surfaces during normal joint function and provides the ability to carry weight and reduce friction. Fluid-film lubrication probably occurs simultaneously in the gaps between the joint surfaces that result from undulations in the articular cartilage surfaces (Fig. 2*e*).

The types of lubrication mechanisms active in a joint depend on the contact pattern of the surfaces, the applied load, and the relative velocity of the surfaces. The contact patterns between joint surfaces vary with different joints and also vary within a joint; depending on the relative positioning of the joint surfaces during a particular motion. The applied load and relative velocity of the surfaces depend on the particular action of the joint. Thus, different lubrication modes may occur within the same joint for different motions or even for the same motion, depending on the applied load and the speed with which the motion is accomplished. *See* SKELETAL SYSTEM.

Van C. Mow; Robert J. Foster

Bibliography. V. C. Mow and W. C. Hayes (eds.), *Basic Orthopaedic Biomechanics*, 2d ed., 1997; V. C. Mow, A. Ratcliffe, and S. L. Y. Woo (eds.), *Biomechanics of Diarthrodial Joints*, vols. 1 and 2, 1990; C. Norkin and P. K. Levangie, *Joint Structure and Function*, 3d ed., 2000.

Joint (structures)

The surface at which two or more mechanical or structural components are united. Whenever parts of a machine or structure are brought together and fastened into position, a joint is formed. *See* STRUCTURAL CONNECTIONS.

Mechanical joints can be fabricated by a great variety of methods, but all can be classified into two general types, temporary (screw, snap, or clamp, for example), and permanent (brazed, welded, or riveted, for example). The following list includes many of the more common methods of forming joints. (1) Screw threads: bolt and nut, machine screw and nut, machine screw into tapped hole, threaded parts (rod, pipe), self-tapping screw, lockscrew, studs with nuts, threaded inserts, coiled-wire inserts, drive screws. (2) Rivets: solid, hollow, explosive and other blind side types. (3) Welding. (4) Soldering. (5) Brazing. (6) Adhesive. (7) Friction-held: nails, dowels, pins, clamps, clips, keys, shrink and press fits. (8) Interlocking: twisted tabs, snap ring, twisted wire, crimp. (9) Other: peening, staking, wiring, stapling, retaining rings, magnetic. Also, pipe joints are made with screw threads, couplings, caulking, and by welding or brazing; masonry joints are made with cement mortar. *See* ADHESIVE; ASPHALT AND ASPHALTITE; BOLT; BOLTED JOINT; BRAZING; MUCILAGE; NUT (ENGINEERING); RIVET; SCREW; SCREW FASTENER; SOLDERING; WASHER; WELDED JOINT; WELDING AND CUTTING OF MATERIALS.

William H. Crouse

Joint disorders

Local or generalized conditions that affect the joints and related tissues. About one person in 16 who is over 15 years old has some form of joint disturbance. The most common conditions are forms of arthritis, which cause the inflammation or degeneration of joint structures. These may follow specific infections, injury, generalized disease states, or degenerative changes which largely parallel the aging processes in the body.

Rheumatoid arthritis, one of the most common diseases which involve joints, is of unknown etiology and occurs most often in young or middle-aged adults. Multiple joints are involved, especially those of the small bones in the hands, feet, and wrists or the jaw. Other connective tissue diseases, such as systemic lupus erythematosus and vasculitis, frequently are accompanied by joint symptomatology.

Degenerative joint disease (osteoarthritis) is a ubiquitous joint disease prevalent in older people. This form of arthritis is characterized by deterioration of the articular surfaces of the weight-bearing joints of the knees, hips, and spine. Frequent involvement is seen at the distal interphalangeal joints

(Heberden's nodes). Arthritis produced by trauma may also result in degenerative joint disease.

Infectious arthritis may be caused by gonorrhea, pneumococci, streptococci, and other pathogenic organisms such as tuberculosis and syphilis.

Gout is an example of crystal-deposition arthritis caused by the presence of monosodium urate crystals in synovial fluid. There is an abnormality of purine metabolism that eventuates in the accumulation of urates and uric acid in joint tissues. The physical findings and clinical course are usually distinctive. *See* GOUT.

Spondyloarthropathies constitute a group of arthritis disorders characterized by involvement of the axial (central) skeleton and association with the histocompatibility system.

Miscellaneous forms of arthritis include those systemic diseases of unknown or quite different etiology which may produce arthritis or joint degeneration. Syringomyelia, syphilis (tabes dorsalis), and leprosy are examples in which destructive joint lesions may be secondary to nerve damage of the affected joints (Charcot joints). Hemophilia and sickle-cell anemia (blood disorders) and acromegaly, hypothyroidism, and hyperparathyroidism (endocrine disorders) are also examples of these unusual rheumatic diseases. *See* ARTHRITIS.

Tumors of joint tissues are usually benign chondromas, fibromas, and lipomas, but malignant synovio sarcoma can occur. *See* TUMOR.

Rheumatism is a nonspecific, predominantly lay term which includes local pain and tenderness. Most causes are chronic inflammations or mildly progressive degenerations, and many, when investigated, fall into one of the previously mentioned categories. *See* BURSITIS; RHEUMATISM.

Robert Searles

Jojoba

Simmondsia chinensis, the only plant known to produce and store a liquid wax in its seed. The jojoba plant is native to the southwestern United States and Mexico. It is tolerant of some of the highest temperatures and most arid regions, and is being domesticated as a crop for hot low-rainfall regions around the world. A broadleaf evergreen shrub that is typically 3–10 ft (1–3 m) in height, it can grow as tall as 20 ft (6 m).

Whereas the seed-storage lipid of other oilseed plants such as soybean and sunflower is a branched ester based on the three-carbon glycerol molecule, that of jojoba is a straight-chain ester. A majority of the wax molecules of jojoba are formed from acids and alcohols with 20 or 22 carbon atoms and one double bond. Many modifications can be made at the double bond, which results in the plant's versatility as an ingredient in a wide range of chemical products. Jojoba wax, used in cosmetics and lubricants, has the potential to serve as a basic feedstock if seed production costs are reduced. Jojoba is a relatively new commercial crop that is being developed simultaneously in many places around the world, and cultivation methods are variable and change rapidly. *See* FAT AND OIL (FOOD); WAX, ANIMAL AND VEGETABLE.

Site and variety selection. Jojoba is a perennial, and flower buds that are formed in the spring, summer, and fall of one year must overwinter before producing a crop in the following year. The buds can succumb at temperatures below −3.5°C (25°F), and so sites where such temperatures are expected are unsuitable for planting. Jojoba is sensitive to standing water, and so sites with well-drained soil away from potential flooding should be selected.

Because widely tested cultivars of jojoba are unavailable, if jojoba is being introduced to an area, small test plantings should first be established and locally adapted varieties selected. Jojoba is dioecious, and both male and female plants are needed for seed production.

Propagation. Jojoba is easy to grow from seed. Germination is best in the temperature range of 68–95°F (20–35°C). Seed-propagated plants are extremely variable in yield, size, sex, and other traits, and so use of seed is not recommended for management systems that require high yield or plant uniformity.

Several asexual methods of propagation are effective for jojoba, including stem cuttage, micropropagation, layering, and grafting. These methods allow propagation of selected plants that have a high yield potential or other desired traits. Stem cuttage is the most widely used. Stem tips with four to six pairs of leaves are collected during the warmer months of the year, basally treated with an auxin, and then rooted by using standard mist propagation techniques. Rooting occurs in 4–8 weeks, and plants are ready for field planting in 4–6 months. *See* REPRODUCTION (PLANT).

Irrigation. Although jojoba grows under natural rainfall conditions in several countries, including Paraguay and Australia, and has grown well in some plantings with natural rainfall in Arizona, most growers irrigate with a drip, furrow, or sprinkler system. Water requirements of jojoba for profitable cultivation are not well defined. In nature, jojoba grows where average annual rainfall is as low as 3 in. (76 mm), but growth is confined to washes, where rainfall is concentrated. The most vigorous plants in nature occur where average annual rainfall is 10–20 in. (250–500 mm). In commercial plantings, good yields have been obtained by using about 1.5 ft (460 mm) of water in the central valley of California and about 2 ft (600 mm) in south-central Arizona. Some growers, however, use 4 ft (1200 mm) or more. Restricting irrigation as winter approaches appears to improve frost resistance. *See* IRRIGATION (AGRICULTURE).

Insects and diseases. Several pest and disease problems have been encountered for which control strategies must be developed. Damage has resulted from attacks by spider mites and several types of insects including stinkbugs, inchworms, thrips, grasshoppers, and lepidopterous larvae. In addition, relatively innocuous infestations with aphids, scales, and mealy bugs have been observed.

Several root and foliar fungal pathogens have been isolated from diseased jojoba plants growing in both field and nursery environments. To date, virus diseases have not been found in association with jojoba. *See* PLANT PATHOLOGY.

Harvesting and processing. Although economic conditions in many countries permit hand harvesting, successful cultivation in the United States will probably rely on an efficient mechanical harvester. Growers have evaluated many types of mechanical harvesters, some of which shake the plants and catch the seed as it falls while others vacuum the seed from the ground after it has dropped. Harvesting requires several passes because of nonuniform maturity. Once collected, the seeds are cleaned by using standard technology developed for other oilseeds. The wax is then pressed or solvent-extracted from the seed. After filtering, a transparent, yellowish oil remains that may undergo an additional decolorization process. David A. Palzkill

Bibliography. H. S. Gentry, The natural history of jojoba (*Simmondsia chinensis*) and its cultural aspects, *Econ. Bot.*, 12:261–295, 1958; G. Robbelen, R. K. Downey, and D. A. Ashri (eds.), *Oil Crops of the World*, 1989; W. C. Sherbrooke and E. F. Haase, *Jojoba: A Wax Producing Shrub of the Sonoran Desert*, 1974; J. Wisniak, *The Chemistry and Technology of Jojoba Oil*, 1987.

Josephson effect

The passage of paired electrons (Cooper pairs) through a weak connection (Josephson junction) between superconductors, as in the tunnel passage of paired electrons through a thin dielectric layer separating two superconductors.

Nature of the effect. Quantum-mechanical tunneling of Cooper pairs through a thin insulating barrier (on the order of a few nanometers thick) between two superconductors was theoretically predicted by Brian D. Josephson in 1962. Josephson calculated the currents that could be expected to flow during such superconductive tunneling, and found that a current of paired electrons (supercurrent) would flow in addition to the usual current that results from the tunneling of single electrons (single or unpaired electrons are present in a superconductor along with bound pairs). Josephson specifically predicted that if the current did not exceed a limiting value (the critical current), there would be no voltage drop across the tunnel barrier. This zero-voltage current flow resulting from the tunneling of Cooper pairs is known as the dc Josephson effect. Josephson also predicted that if a constant nonzero voltage V were maintained across the tunnel barrier, an alternating supercurrent would flow through the barrier in addition to the direct current produced by the tunneling of unpaired electrons. The frequency ν of the alternating supercurrent is given by Eq. (1), where e is

$$\nu = \frac{2eV}{h} \tag{1}$$

the magnitude of the charge of an electron and h is Planck's constant. The oscillating current of Cooper pairs that flows when a steady voltage is maintained across a tunnel barrier is known as the ac Josephson effect. Josephson further predicted that if an alternating voltage at frequency f were superimposed on the steady voltage applied across the barrier, the alternating supercurrent would be frequency-modulated and could have a dc component whenever ν was an integral multiple of f. Depending upon the amplitude and frequency of the ac voltage, the dc current-voltage characteristic would display zero-resistance parts (constant-voltage steps) at voltages V given by Eq. (2), where n is any integer. Finally, Josephson

$$V = \frac{nhf}{2e} \tag{2}$$

predicted that effects similar to the above would also occur for two superconducting metals separated by a thin layer of nonsuperconducting (normal) metal. In 1963 the existence of the dc Josephson effect was experimentally confirmed by P. W. Anderson and J. M. Rowell, and the existence of the ac Josephson effect was experimentally confirmed by S. Shapiro. *See* NONRELATIVISTIC QUANTUM THEORY; TUNNELING IN SOLIDS.

Theory of the effect. The superconducting state has been described as a manifestation of quantum mechanics on a macroscopic scale, and the Josephson effect is best explained in terms of phase, a basic concept in the mathematics of quantum mechanics and wave motion. For example, two sine waves of the same wavelength λ are said to have the same phase if their maxima coincide, and to have a phase difference equal to $2\pi/\lambda$ if their maxima are displaced by a distance δ. An appreciation of the importance that phase can have in physical systems can be gained by considering the radiation from excited atoms in a ruby rod. For a given transition, the atoms emit radiation of the same wavelength; if the atoms also emit the radiation in phase, the result is the ruby laser.

According to the Bardeen-Cooper-Schrieffer (BCS) theory of superconductivity, an electron can be attracted by the deformation of the metal lattice produced by another electron, and thereby be indirectly attracted to the other electron. This indirect attraction tends to unite pairs of electrons having equal and opposite momentum and antiparallel spins into the bound pairs known as Cooper pairs. In the quantum-mechanical description of a superconductor, all Cooper pairs in the superconductor have the same wavelength and phase. It is this phase coherence that is responsible for the remarkable properties of the superconducting state. The common phase of the Cooper pairs in a superconductor is referred to simply as the phase of the superconductor. *See* COHERENCE; PHASE (PERIODIC PHENOMENA); QUANTUM MECHANICS.

The phases of two isolated superconductors are totally unrelated, while two superconductors in perfect contact have the same phase. If the

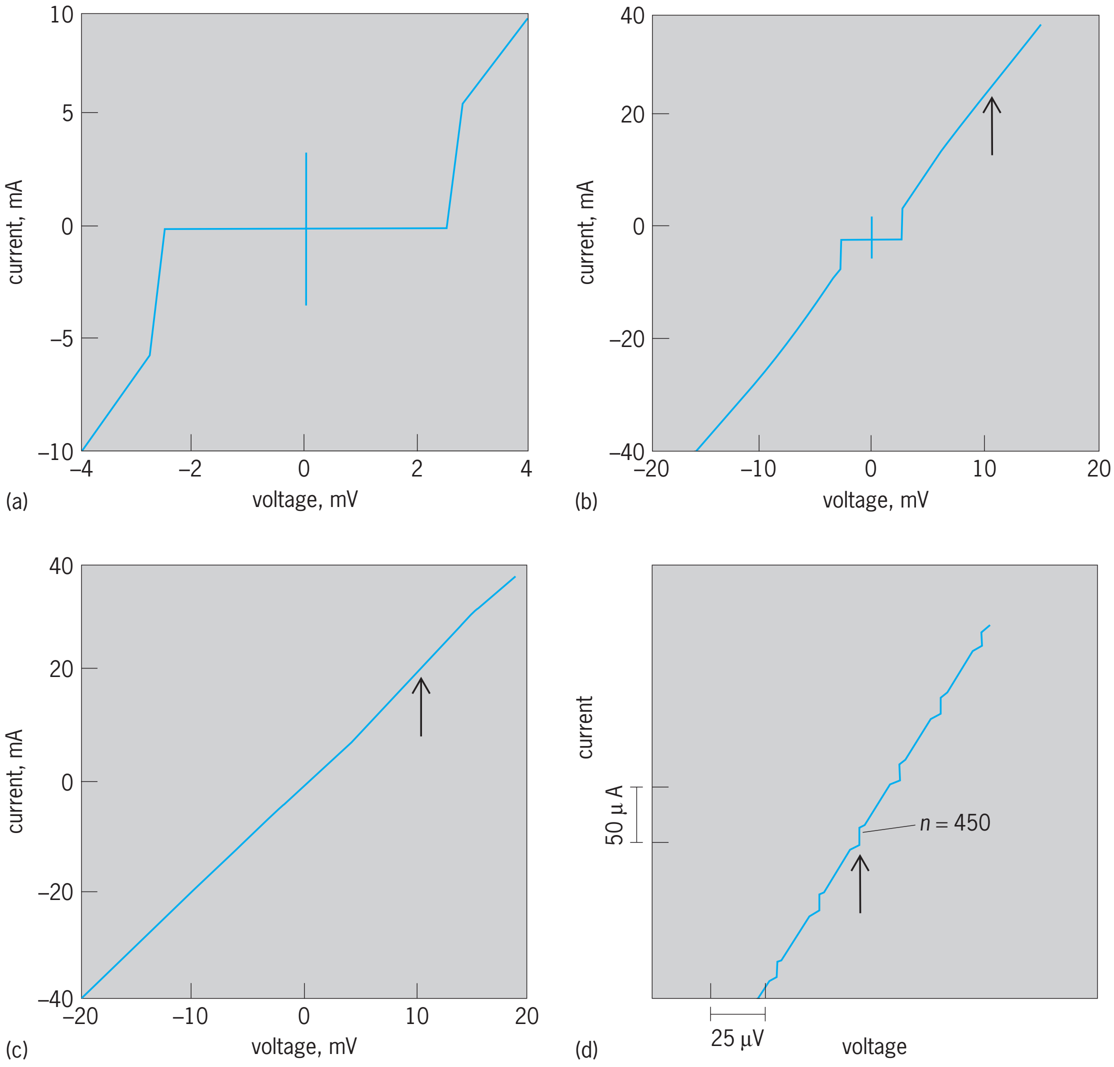


Fig. 1. DC current-voltage characteristics of lead-lead oxide-lead Josephson tunnel junction at 1.2 K. (*a*) Without microwave power. (*b*) Same characteristic with reduced scale. (*c*) 11-GHz microwave power applied. (*d*) Expanded portion of *c*; arrow indicates a constant-voltage step near 10.2 mV corresponding to $n = 450$ in Eq. (2). This voltage is also indicated by arrows in *b* and *c*. (*After T. F. Finnegan, A. Denenstein, and D. N. Langenberg, AC-Josephson-effect determination of e/h: A standard of electrochemical potential based on macroscopic quantum phase coherence in superconductors, Phys. Rev. B, 4:1487–1522, 1971*)

superconductors are weakly connected (as they are when separated by a sufficiently thin tunnel barrier), the phases can be different but not independent. If ϕ is the difference in phase between superconductors on opposite sides of a tunnel barrier, the results of Josephson's calculation of the total current I through the junction can be written as Eq. (3), where I_0 is the

$$I = I_0 + I_1 \sin\phi \tag{3}$$

current due to single electron tunneling, and $I_1 \sin\phi$ is the current due to pair tunneling. The time dependence of the phase is given by Eq. (4). In general, the

$$\frac{\partial\phi}{\partial t} = 2\pi\frac{2eV}{h} \tag{4}$$

currents I, I_0, and I_1 are all functions of the voltage across the junction. For $V = 0$, I_0 is zero and ϕ is constant. The value of I_1 depends on the properties of the tunnel barrier, and the zero-voltage supercurrent is a sinusoidal function of the phase difference between the two superconductors. However, it is not the phase difference that is under the control of the experimenter, but the current through the junction, and the phase difference adjusts to accommodate the current. The maximum value sin ϕ can assume is 1, and so the zero-voltage value of I_1 is the critical current of the junction.

Integration of Eq. (4) shows the phase changes linearly in time for a constant voltage V maintained across the barrier, and the current through the barrier is given by Eq. (5), where ϕ_0 is a constant. The

$$I = I_0 + I_1 \sin\left(2\pi\frac{2eV}{h}t + \phi_0\right) \tag{5}$$

supercurrent is seen to be an alternating current with frequency $2eV/h$. The supercurrent time-averages to zero, and so the direct current through the barrier is just the single-electron tunneling current I_0.

If the voltage across the junction is $V + v\cos 2\pi f t$, Eqs. (3) and (4) give Eq. (6) for the current. The expression for the supercurrent is a conventional expression in frequency-modulation theory and can be rewritten as expression (7), where J_n is an integer order Bessel function of the first kind.

$$I = I_o + I_1 \sin\left(2\pi \frac{2eV}{h}\, t + \phi_0 + \frac{2ev}{hf}\sin 2\pi f t\right) \qquad (6)$$

$$I_1 \sum_{n=-\infty}^{n=\infty} +(-1)^n J_n \frac{2ev}{hf} \times \sin\left(2\pi \frac{2eV}{h} t - 2\pi n f t + \phi_0\right) \qquad (7)$$

This expression time-averages to zero except when $V = nhf/2e$, in which case the supercurrent has a dc component given by $(-1)^n I_1 J_n(2ev/hf)\sin\phi_0$. As for the zero-voltage direct supercurrent, the phase difference ϕ_0 adjusts to accommodate changes in current at this value of V, and the dc current-voltage characteristic displays a constant-voltage step.

The dc current-voltage characteristic of a Josephson tunnel junction with and without a microwave-frequency ac voltage is shown in **Fig. 1**. The straightening of the current-voltage characteristic in the presence of microwave power displayed in Fig. 1*c* is due to the phenomenon of photon-assisted tunneling, which is essentially identical to classical rectification for the junction and frequency in question. *See* BESSEL FUNCTIONS.

Josephson pointed out that the magnitude of the maximum zero-voltage supercurrent would be reduced by a magnetic field. In fact, the magnetic field dependence of the magnitude of the critical current is one of the more striking features of the Josephson effect. Circulating supercurrents flow through the tunnel barrier to screen an applied magnetic field from the interior of the Josephson junction just as if the tunnel barrier itself were weakly superconducting. The screening effect produces a spatial variation of the transport current, and the critical current goes through a series of maxima and minima as the field is increased. **Figure 2** shows the variation of the critical current with magnetic field for a tunnel junction whose length and width are small in comparison with the characteristic screening length of the junction (the Josephson penetration depth, λ_J). The mathematical function which describes the magnetic field dependence of the critical current for this case is the same function as that which describes the diffraction pattern produced when light passes through a single narrow slit. *See* DIFFRACTION.

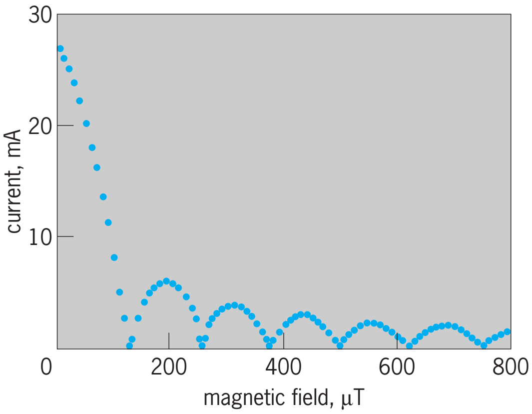


Fig. 2. Magnetic field dependence of the critical current of a Josephson tunnel junction. Data are for a tin-tin oxide-tin junction at 1.2 K, with the magnetic field in the plane of the barrier. (*After D. N. Langenberg, D. J. Scalapino, and B. N. Taylor, The Josephson effects, Sci. Amer., 214(5):30–39, May 1966*)

Josephson junctions. The weak connections between superconductors through which the Josephson effects are realized are known as Josephson junctions. Historically, superconductor-insulator-superconductor tunnel junctions have been used to study the Josephson effect, primarily because these are physical situations for which detailed calculations can be made. However, the Josephson effect is not necessarily a tunneling phenomenon, and the Josephson effect is indeed observed in other types of junctions, such as the superconductor-normal metal-superconductor junction. A particularly useful Josephson junction, the point contact, is formed by bringing a sharply pointed superconductor into contact with a blunt superconductor. The critical current of a point contact can be adjusted by changing the pressure of the contact. The low capacitance of the device makes it well suited for high-frequency applications. Thin-film microbridges form another group of Josephson junctions. The simplest microbridge is a short narrow constriction (length and width on the order of a few micrometers or smaller) in a superconducting film known as the Anderson-Dayem bridge. If the microbridge region is also thinner than the rest of the superconducting film, the resulting variable-thickness microbridge has better performance in most device applications. If a narrow strip of superconducting film is overcoated along a few micrometers of its length with a normal metal, superconductivity is weakened beneath the normal metal, and the resulting microbridge is known as a proximity-effect or Notarys-Mercereau microbridge. Among the many other types of Josephson junctions are the superconductor-semiconductor-superconductor and other artificial-barrier tunnel junctions, superconductor-oxide-normal metal-superconductor junctions, and the so-called SLUG junction, which consists of a drop of lead-tin solder solidified around a niobium wire. Some different types of Josephson junctions are illustrated in **Fig. 3**.

The dc current-voltage characteristics of different types of Josephson junctions may differ, but all show a zero-voltage supercurrent, and constant-voltage steps can be induced in the dc characteristics at voltages given by Eq. (2) when an ac voltage is applied. The dc characteristics of a microbridge and a tunnel junction are compared in **Fig. 4**.

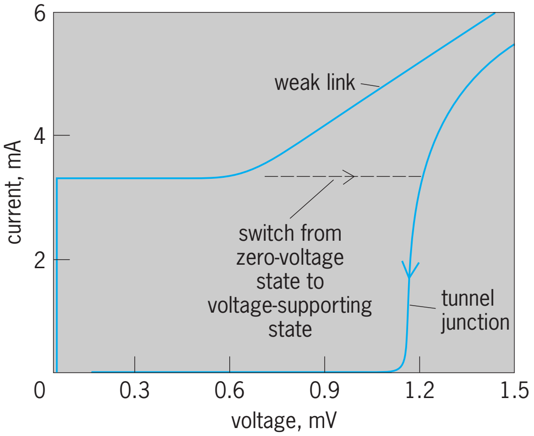


Fig. 4. DC current-voltage characteristics for a weak link and a tunnel junction. (*After E. Burstein and S. Lundqvist, eds., Tunneling Phenomena in Solids, Plenum, 1969*)

Applications. From July 1972 through December 1989, the United States legal volt, V_{NBS}, was defined by Eq. (1) through the assigned value given by Eq. (8),

$$\frac{2e}{h} = 483,593,420\ \text{MHz}/\text{V}_{\text{NBS}} \tag{8}$$

and was maintained at the National Institute of Standards and Technology (formerly the National Bureau of Standards) within a few parts in 10^8 by using the ac Josephson effect. By international agreement, a new practical reference standard for the volt, based on the ac Josephson effect, was adopted worldwide on January 1, 1990. This internationally adopted representation of the volt is given by Eq. (9), where K_J

$$K_{J-90} = 483,597.9\ \text{GHz/V} \tag{9}$$

is called the Josephson constant and is theoretically equal to $2e/h$. This developed as a natural consequence of extremely precise measurements of $2e/h$ via the Josephson effect, and the recognition that a Josephson junction is a precise frequency-to-voltage converter and that atomic frequency standards are inherently more stable than electrochemical voltage standards. *See* ELECTRICAL UNITS AND STANDARDS; FUNDAMENTAL CONSTANTS.

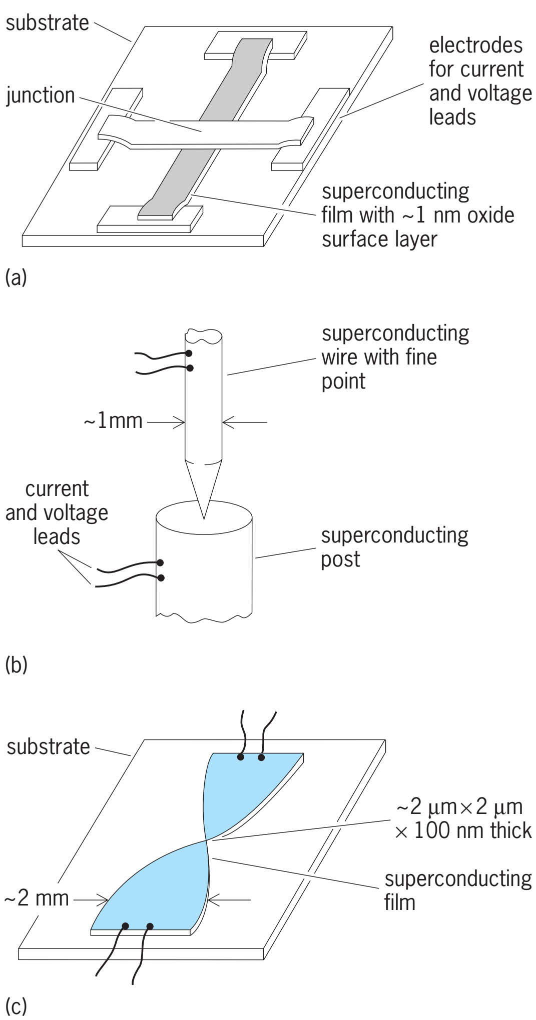


Fig. 3. Some types of Josephson junctions. (*a*) Thin-film tunnel junction. (*b*) Point contact. (*c*) Thin-film weak link. (*After E. Burstein and S. Lundqvist, eds., Tunneling Phenomena in Solids, Plenum, 1969*)

The Josephson effect permits measurement of absolute temperature: a voltage drop across a resistor in parallel with a Josephson junction causes the junction to emit radiation at the frequency given by Eq. (1), but voltage fluctuations resulting from thermal noise produce frequency fluctuations that depend on absolute temperature. *See* LOW-TEMPERATURE THERMOMETRY.

Josephson junctions, and instruments incorporating Josephson junctions, are used in other applications for metrology at dc and microwave frequencies, frequency metrology, magnetometry, detection and amplification of electromagnetic signals, and other superconducting electronics such as high-speed analog-to-digital converters and computers. A Josephson junction, like a vacuum tube or a transistor, is capable of switching signals from one circuit to another; a Josephson tunnel junction is capable of switching states in as little as 6 picoseconds and is the fastest switch known. Josephson junction circuits are capable of storing information. Finally, because a Josephson junction is a superconducting device, its power dissipation is extremely small, so that Josephson junction circuits can be packed together as tightly as fabrication techniques will permit. All the basic circuit elements required for a Josephson junction computer have been developed. *See* SUPERCONDUCTING DEVICES; SUPERCONDUCTIVITY.

Louis B. Holdeman

Bibliography. A. Barone and G. Paterno, *Physics and Applications of the Josephson Effect*, 1982; A. De Chiara and M. Russo, *Tunneling Phenomena in High and Low T_c Superconductors*, 1993.

Joule's law

A quantitative relationship between the quantity of heat produced in a conductor and an electric current flowing through it. As experimentally determined and announced by J. P. Joule, the law states that when a current of voltaic electricity is propagated along a metallic conductor, the heat evolved in a given time is proportional to the resistance of the conductor multiplied by the square of the electric intensity. Today the law would be stated as $H = RI^2$, where H is rate of evolution of heat in watts, the unit of heat being the joule; R is resistance in ohms; and I is current in amperes. This statement is more general than the one sometimes given that specifies that R be independent of I. Also, it is now known that the application of the law is not limited to metallic conductors.

Although Joule's discovery of the law was based on experimental work, it can be deduced rather easily for the special case of steady conditions of current and temperature. As a current flows through a conductor, one would expect the observed heat output to be accompanied by a loss in potential energy of the moving charges that constitute the current. This loss would result in a descending potential gradient along the conductor in the direction of the current flow, as usually defined. If E is the total potential drop, this loss, by definition, is equal to E in joules for every coulomb of charge that traverses the conductor. The loss conceivably might appear as heat, as a change in the internal energy of the conductor, as work done on the environment, or as some combination of these. The second is ruled out, however, because the temperature is constant and no physical or chemical change in a conductor as a result of current flow has ever been detected. The third is ruled out by hypothesis, leaving only the generation of heat. Therefore, $H = EI$ in joules per second, or watts. By definition, $R = E/I$, a ratio which has positive varying values. Elimination of E between these two equations gives the equation below, which is Joule's law as stated

$$H = RI^2$$

above. If I changes to a new steady value I', R to R', and H and $H' = R'I'^2$ as before. The simplest case occurs where R is independent of I. If the current is varying, the resulting variations in temperature and internal energy undoubtedly exist and, strictly speaking, should be allowed for in the theory. Yet, in all but the most exceptional cases, any correction would be negligible.

This phenomenon is irreversible in the sense that a reversal of the current will not reverse the outflow of heat, a feature of paramount importance in many problems in physics and engineering. Thus the heat evolved by an alternating current is found by taking the time average of both sides of the equation. Incidentally, the changes in internal energy, if they were included in the theory, would average out. Hence the equation continues to have a similar form, $\overline{H} = \overline{RI^2}$, for ac applications. *See* ELECTRIC HEATING; OHM'S LAW.

Lewellyn G. Hoxton; John W. Stewart

Juglandales

An order of flowering plants, division Magnoliophyta (Angiospermae), in the subclass Hamamelidae of the class Magnoliopsida (dicotyledons). The order consists of two families: the Juglandaceae with a little over 50 species and the Rhoipteleaceae with only one species. Within its subclass the order is sharply set off by its compound leaves. *Juglans* (walnut and butternut) and *Carya* (hickory, including the pecan, *C. illinoensis*) are familiar genera of the Juglandaceae. *See* HAMAMELIDAE; HICKORY; MAGNOLIOPHYTA; MAGNOLIOPSIDA; PLANT KINGDOM.

Arthur Cronquist; T. M. Barkley

Juncales

An order of flowering plants, division Magnoliophyta (Angiospermae), in the subclass Commelinidae of the class Liliopsida (monocotyledons). The order consists of the family Juncaceae, with about 300 species, and the family Thurniaceae, with only three. Within its subclass the order is marked by its reduced, mostly wind-pollinated flowers and capsular fruits with one to many anatropous ovules per carpel. The flowers have six sepals arranged in two more or less similar whorls, both sets chaffy and usually brown or green. The ovary is tricarpellate, with axile or parietal placentation. The pollen grains are borne in tetrads, and the embryo is surrounded by endosperm. The order is most unusual among higher plants in that, together with at least some members of the Cyperaceae in the related order Cyperales, it has chromosomes with diffuse centromeres. *See* COMMELINIDAE; CYPERALES; LILIOPSIDA; MAGNOLIOPHYTA; PLANT KINGDOM.

Arthur Cronquist

Junction detector

A device in which detection of radiation takes place in or near the depletion region of a reverse-biased semiconductor junction. The electrical output pulse is linearly proportional to the energy deposited in the junction depletion layer by the incident ionizing radiation. *See* CRYSTAL COUNTER; IONIZATION CHAMBER.

Introduced into nuclear studies in 1958, the junction detector, or more generally, the nuclear semiconductor detector, revolutionized the field. In the detection of both charged particles and gamma radiation, these devices typically improved experimentally attainable energy resolutions by about two

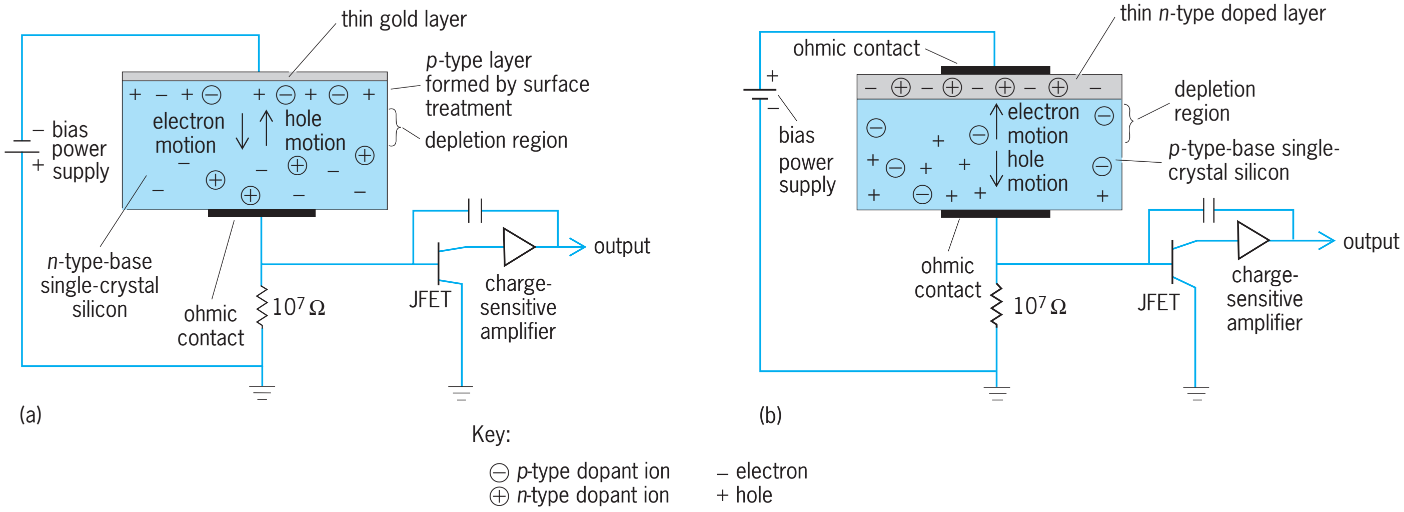


Fig. 1. Silicon junction detectors. (*a*) Surface barrier detector. (*b*) A *pn* junction detector. The *p*-type dopant ions are fixed in the crystal lattice. JFET = junction field-effect transistor.

orders of magnitude over that previously attainable. To this they added unprecedented flexibility of utilization, speed of response, miniaturization, freedom from deleterious effects of extraneous electromagnetic (and often nuclear) radiation fields, low-voltage requirements, and effectively perfect linearity of output response. They are now used for a wide variety of diverse applications, from logging drill holes for uranium to examining the Shroud of Turin. They are used for general analytical applications, giving both qualitative and quantitative analysis in the microprobe and the scanning transmission electron microscopes. They are used in medicine, biology, environmental studies, and the space program. In the last category they play a very fundamental role, ranging from studies of the radiation fields in the solar system to the composition of extraterrestrial surfaces. *See* ELECTRON MICROSCOPE; SECONDARY ION MASS SPECTROMETRY (SIMS).

Fabrication of diodes. The first practical detectors were prepared by evaporating a very thin gold layer on a polished and etched wafer of *n*-type germanium (Ge). To reduce noise these devices were operated at liquid nitrogen temperature (77 K). Silicon (Si), however, with its larger band gap, 1.107 eV compared to 0.67 eV for germanium, offered the possibility of room-temperature operation. Gold-silicon surface barrier detectors and silicon *pn* junction detectors were soon developed.

Surface barrier detectors are made from wafers of *n*-type silicon semiconductor crystals. The etching and surface treatments create a thin *p* layer, and the gold contacts this layer (**Fig. 1***a*). The *pn* junction silicon detectors are usually made by diffusing phosphorus about 2 micrometers into the surface of a *p*-type silicon base (Fig. 1*b*). Both techniques give a *pn* junction. When this junction is reverse-biased, a depletion region, or a region devoid of carriers (electrons and holes), forms mainly in the higher-resistivity base material. A high field now exists in this region, and any carriers born or generated in it are rapidly swept from the region. The requirement for detection is that the ionizing radiation must lose its energy by creating electron-hole pairs (2.96 eV/pair in germanium and 3.66 eV/pair in silicon) in the depletion region or within a carrier diffusion length of this region. Both carriers have to be collected to give an output pulse proportional to the energy of the incident particle. Electrons and holes have similar mobilities in both silicon and germanium, and although carrier trapping occurs it is not as severe as in the 2–16 compounds.

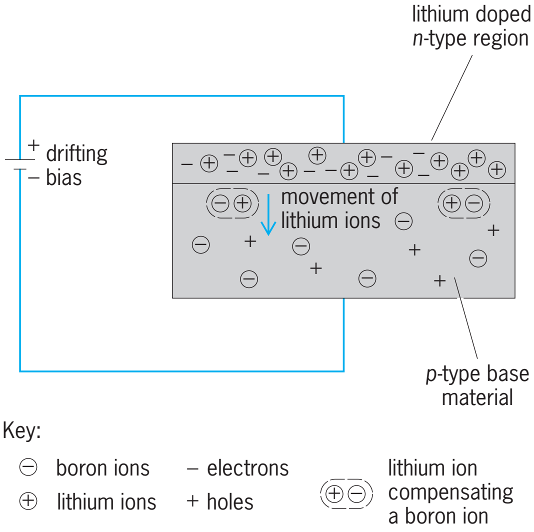


Fig. 2. Compensation of *p*-type semiconductor material with lithium (at 212–392°F or 100–200°C for silicon, 104–140°F or 40–60°C for germanium). Boron ions are fixed in the lattice. Lithium ions are fixed in the lattice, but at elevated temperature can be drifted under an electric field and will compensate boron ions to widen the depletion region.

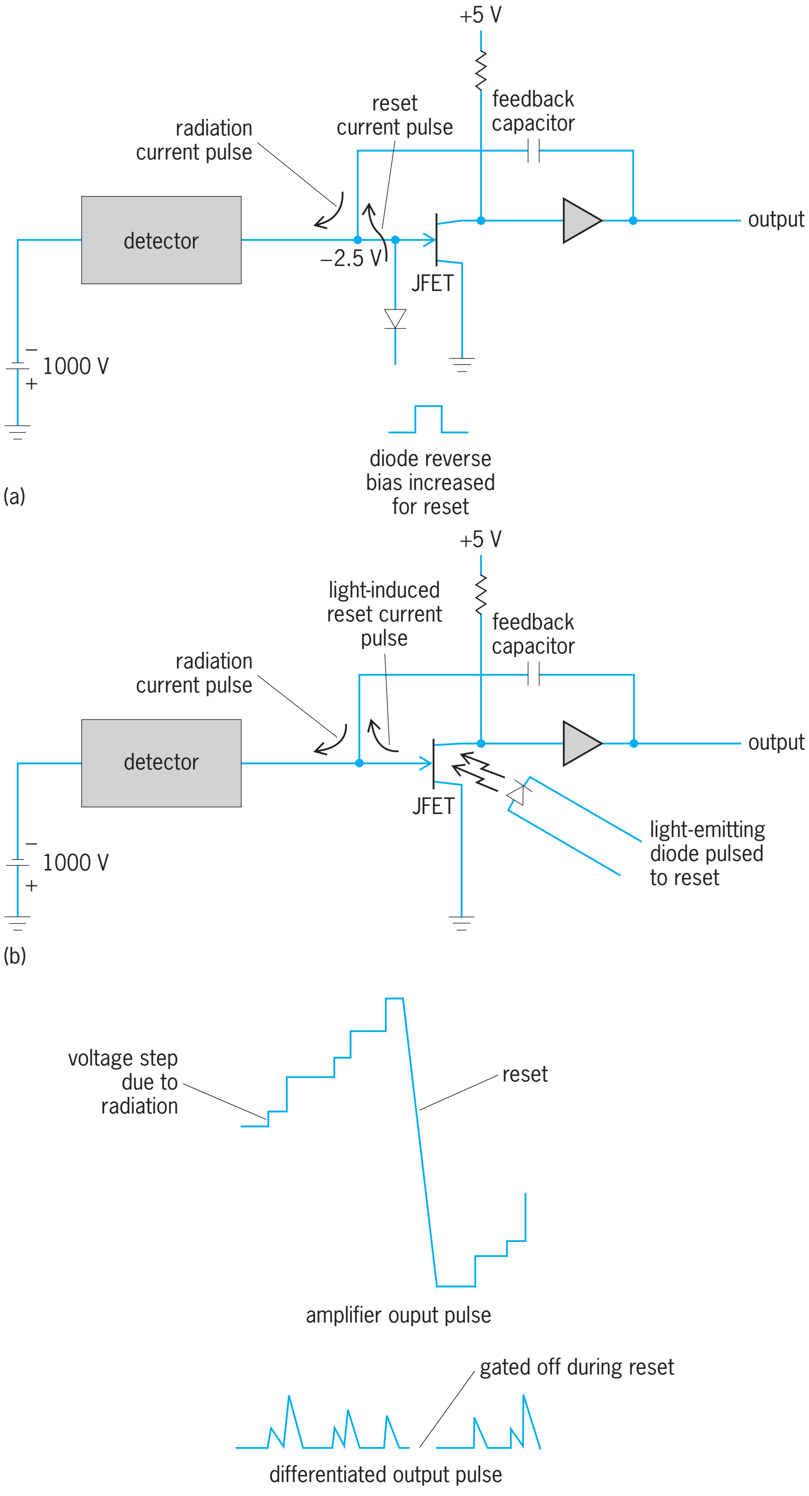


Fig. 3. Reset mechanisms for junction detectors. (*a*) Diode reset. (*b*) Optical reset. (*c*) Amplifier output and differentiated output for the pulse height analyzer.

Control of depletion region width. The detection of charged particles in the presence of gamma rays or higher-energy particles can be optimized by controlling the width of the depletion region. This width is a function of the reverse bias and of the resistivity of the base material. There is a practical limit to the voltage that can be applied to a junction. Thus detectors for higher-energy or lower-mass particles (electrons) requiring wider depletion regions are made from high-resistivity material. This material occurs, by accident, during the growth of some crystals.

Lithium-drifted silicon detectors. Still wider depletion-width detectors can be made from lithium-drifted silicon. Lithium (Li) is a donor in silicon. In addition, at elevated temperatures (392°F or 200°C), the lithium ion is itself mobile. Thus when lithium is diffused in *p*-type silicon, a *pn* junction results. Reverse-biasing this junction at elevated temperatures causes the lithium ion, now appearing as a positive charge, to migrate toward the negative side. On the way it encounters an acceptor ion, negatively charged, which is fixed in the crystal lattice. The lithium ion and the acceptor ion compensate each other, and the lithium ion remains in this location. As more lithium ions drift into and across the depletion region, they compensate the acceptor ions and the region widens (**Fig. 2**). Depletion regions, or compensated regions, up to 0.8 in. (2 cm) wide have been achieved with this technique. Lithium-drifted silicon detectors can be operated at room temperature, but the larger volume gives a greater thermally generated leakage current, which degrades the resolution. The best energy resolution is obtained by operating the detectors at low temperature. However, they may be stored at room temperature.

Lithium-drifted silicon detectors are widely used to detect particle- or photon-induced x-rays. The resolution, when operated at 77 K, is sufficient to resolve the K x-rays for all elements higher in atomic number Z than carbon ($Z = 6$). A resolution of 100 eV has been obtained at 2 keV. At the lower x-ray energies the effects of the detector window thickness and the absorption in the window of the mounting are important, and silicon is preferred for these applications. For x-rays the efficiency of a 5-mm depletion-width lithium-drifted silicon detector is about 50% at 30 keV and 5% at 60 keV. Typically these detectors have capacitances of about 2 picofarads and, to minimize noise, are operated with an optical or diode reset mechanism rather than a feedback resistor (**Fig. 3**). The detector bias is about 1000 V, and the junction field-effect transistor (JFET) gate operates at about −2.5 V. A radiation event causes a pulse of current in the detector. The amplifier drives this current i through the feedback capacitor with capacitance C and in doing so steps a voltage an amount e_{step} proportional to the charge, as given by the equation below. Each subsequent radiation event causes a voltage step. To keep the amplifier within its dynamic range the feedback capacitor must be discharged. The analyzing circuits are first gated off, and in the diode case (Fig. 3*a*) the reverse bias on the diode is momentarily increased to give a picoampere current pulse. The amplifier output

$$e_{\mathrm{step}} = \frac{1}{C}\int i\,dt$$

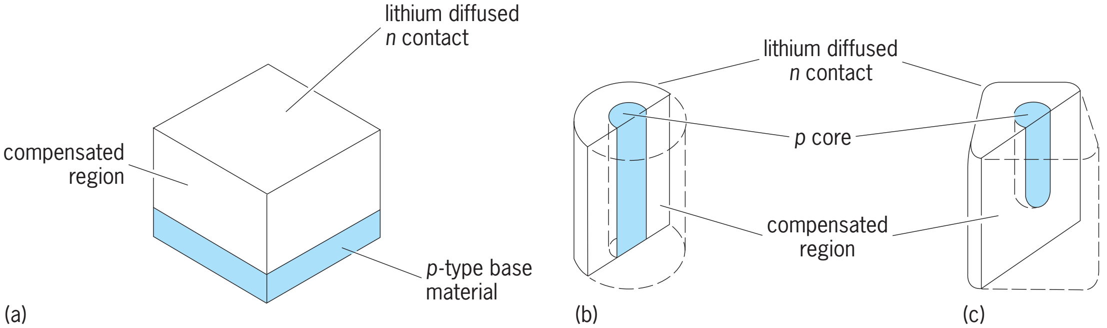


Fig. 4. Lithium-drifted detectors. (*a*) Planar. (*b*) Coaxial. (*c*) Open-one-end coaxial.

voltage changes to allow this current to flow through the feedback capacitor, discharging it. The analyzing circuit is now gated on, and counting can resume. For the optical reset (Fig. 3*b*), a light is flashed on the JFET, momentarily increasing the source-to-gate leakage current and discharging the feedback capacitor. The output from the amplifier and the differentiated output for the analyzer are shown in Fig. 3*c*.

Lithium-drifted germanium detectors. Germanium with its higher atomic number, 32 compared with 14 for silicon, has higher radiation absorption than silicon. Lithium may also be drifted in germanium. But in germanium, lithium is mobile at room temperature and will precipitate or diffuse further if the units, after fabrication, are not kept at liquid nitrogen temperature. Lithium-drifted germanium detectors revolutionized the field of gamma-ray spectroscopy. They may be manufactured in planar, coaxial, or open-one-end coaxial geometry (**Fig. 4**).

Figure 5 compares the gamma-ray spectrum of ^{188}Os taken with a 21-cm^3 lithium-germanium detector with that from a sodium iodide (NaI) scintillator-type spectrometer which is 3 in. (7.5 cm) in diameter by 3 in. deep (27 in.^{3} or 330 cm^3). The counting efficiency of the lithium-germanium detector is lower than the scintillator, but the resolution is at least an order of magnitude better. This higher resolution often reduces the actual counting time to adequately identify a particular energy peak even with an order-of-magnitude less sensitive volume. Also, as shown in Fig. 5, the lithium-germanium detector is able to resolve more energy groups than the scintillator.

Hyperpure germanium detectors. Intrinsic or hyperpure germanium (**Fig. 6**) was grown to overcome the low-temperature-storage and the lithium-drifting problems associated with lithium-germanium. Planar detectors with up to an 0.8-in.-thick (2-cm) depletion region and coaxial detectors with 3-in.^{3} (50-cm^3)

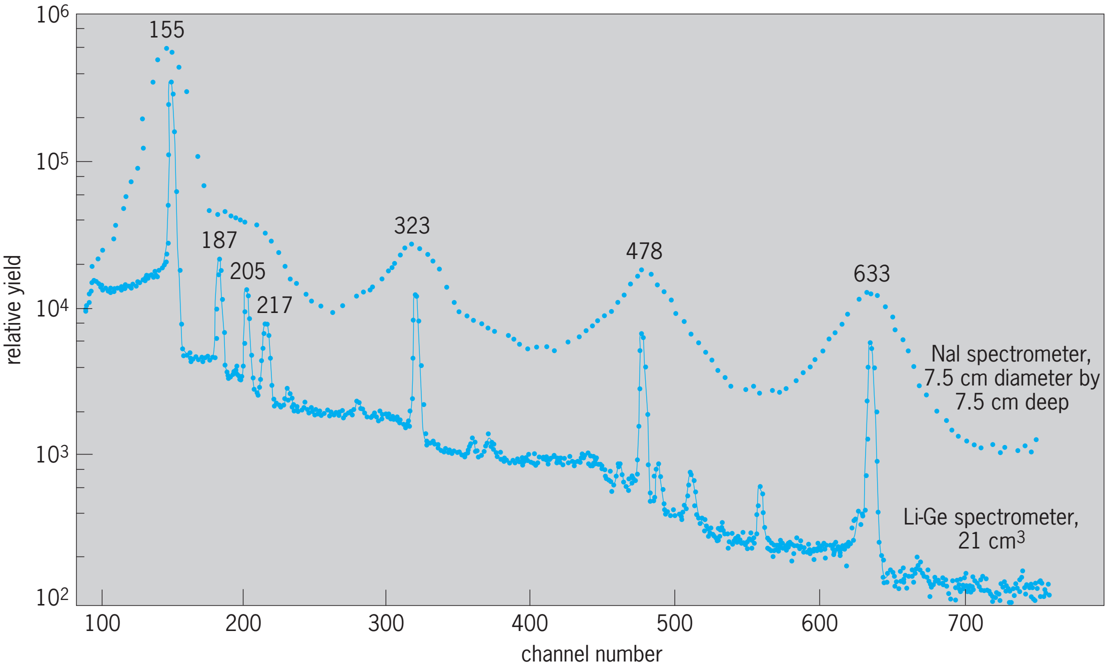


Fig. 5. Gamma-radiation spectra from ^{188}Os as detected in sodium iodide (NaI) and lithium-germanium (Li-Ge) spectrometers. 1 cm = 0.4 in.; 1 cm^3 = 6 × 10^{-2} in.^{3}

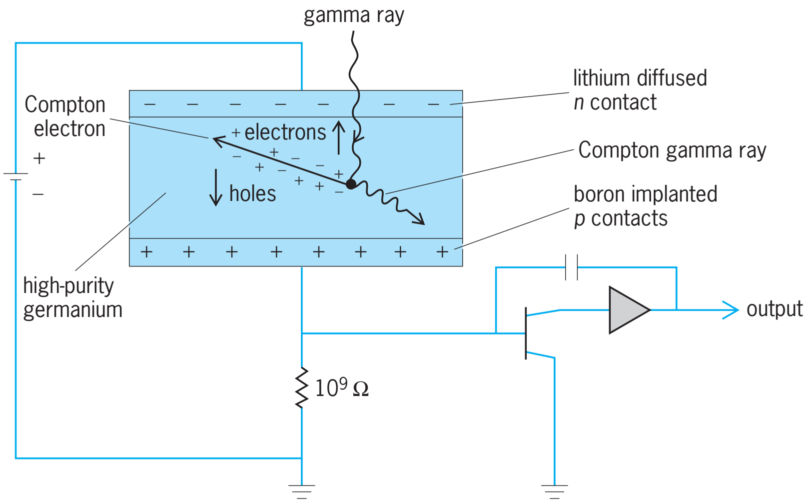


Fig. 6. High-purity germanium gamma-ray detector.

volume have been made with the material. Low-temperature processing is used in the fabrication—usually lithium diffused at 536°F (280°C) for the $n+$ contact and implanted boron for the $p+$ contract. This low-temperature processing is desirable to prevent diffusion of copper, with its subsequent charge trapping, into the germanium. Presently hyperpure germanium detectors cannot be made either as large as, nor with as high a resolution as, lithium-germanium detectors. Both types are operated at liquid nitrogen temperature, 77 K. However, the hyperpure germanium detector is easier to manufacture and can be stored at room temperature when not in use. This is a tremendous practical advantage.

Special detector configurations. Among the many other advantages of semiconductor detectors is the ease with which special detector configurations may be fabricated. One of the simple yet very important examples of this is the annular detector (**Fig. 7**), which is characteristically used to detect nuclear reaction products from a bombarded target in a tight cone around the incident beam. By examining the decay radiation in coincidence with such products, studies may be carried out only on residual nuclei which have had their spins very highly aligned in the nuclear reaction; this has been shown to provide an extremely powerful nuclear spectroscopic probe. The annular detector is extensively used in laboratories worldwide.

Composite detector systems are very readily assembled with the semiconductor devices. For example, it is standard in charged-particle detection to use a very thin detector and a very thick detector (or even two thin and one thick) in series. Multiplication of the resultant signals readily provides a characteristic identification signature for each nuclear particle species in addition to its energy. Three-crystal gamma-ray spectrometers are readily assembled, wherein only the output of the central detector is examined whenever it occurs in time coincidence with two correlated annihilation quanta escaping from the central detector. These systems essentially eliminate background from Compton scattering of other more complex electro-magnetic interactions and yield sharp single peaks for each incident photon energy (**Fig. 8**).

Similarly neutrons may be indirectly detected through examination of recoil protons from a hydrogenous radiator in the case of high-energy neutrons, or through examination of fission fragments resulting from slow neutrons incident on a fissile converter foil mounted with the semiconductor detectors. (It should be noted that the response of the detectors is essentially perfectly linear all the way from electrons and photons to fission fragments.) Neutrons also may be detected and their energy spectra studied through examination of the charged products of the $(n\alpha)$ reaction (where alpha particles are emitted from incident neutrons) induced in the silicon or germanium base material of the detector itself.

Fabrication of triodes. Whereas the detectors thus far discussed are electrically nothing more than diodes, it has been possible to construct equivalent triodes which have extremely important uses in that they provide not only an output which is linearly proportional to the energy deposited in them, but also a second output which in combination with the first establishes the precise location on the detector itself where the ionizing radiation was incident. This has very obvious advantages in the construction of simple systems for the measurement of angular distributions, where such position-sensitive detectors are located about a bombarded target. Their most important impact, however, has been in terms of their on-line use in the focal planes of large nuclear magnetic spectrographs. Simultaneous determination of the energy and location of a particle in the focal plane, together with the momentum

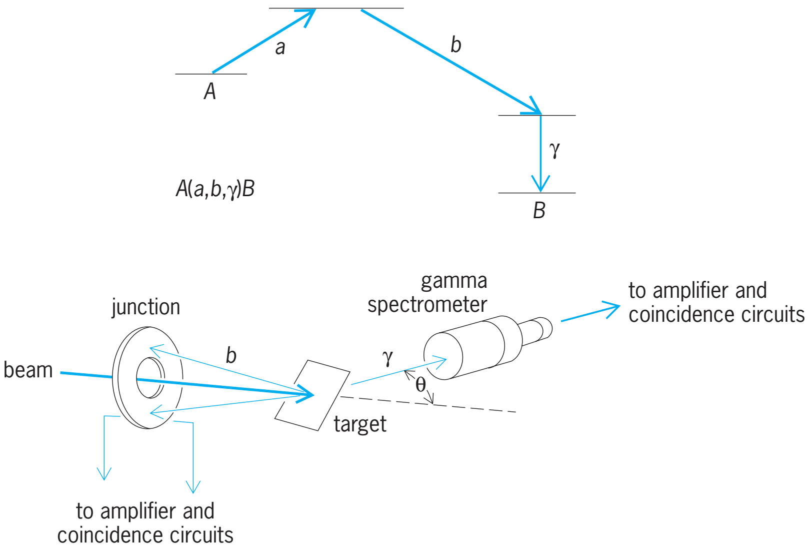


Fig. 7. Schematic of use of annular detector in nuclear reaction studies.

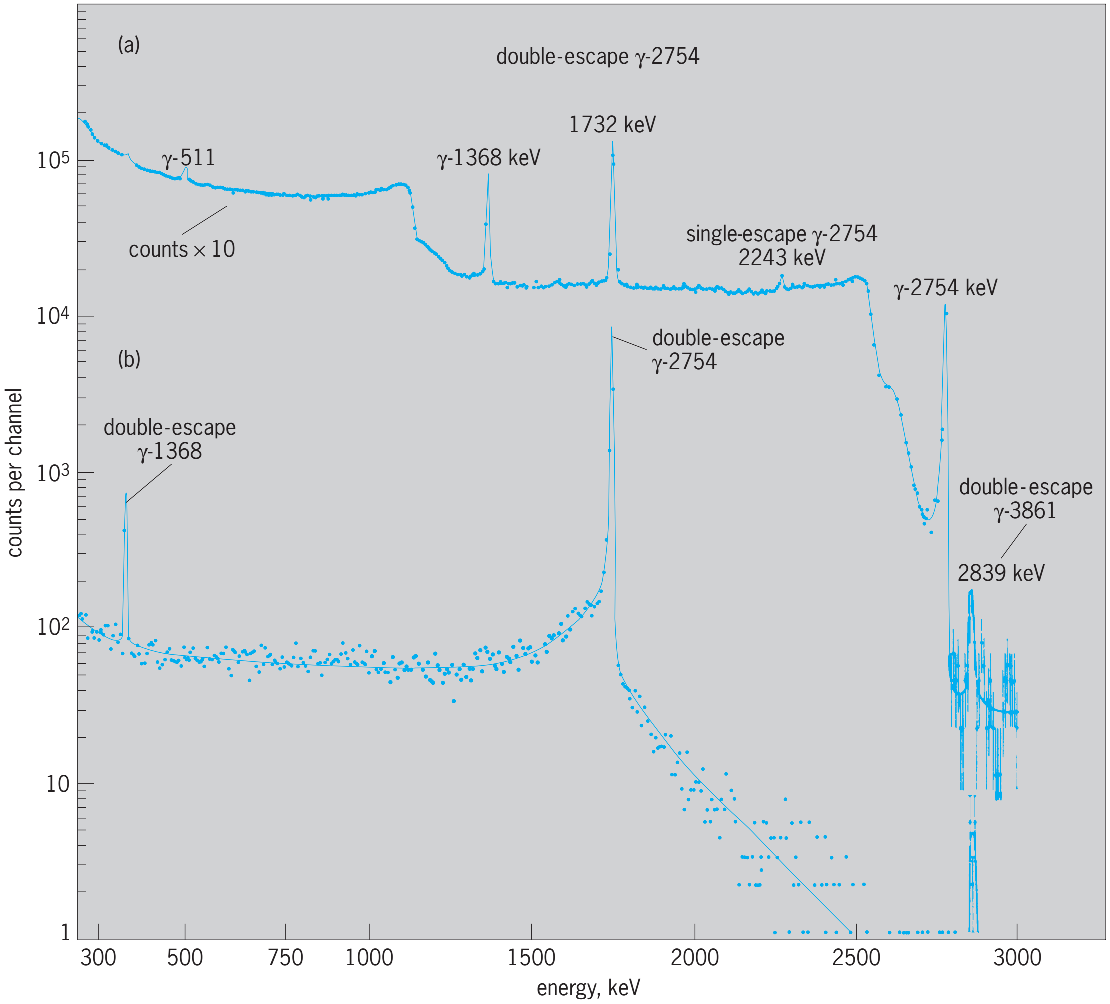


Fig. 8. Comparison of (*a*) direct single detector and (*b*) three-crystal spectrometer spectra from ^{24}Na source.

determination by the magnet itself, establishes unambiguously both the mass and energy of the particle, and does so instantaneously so that additional logical constraints may be imposed through a connected on-line computer—something totally impossible with the earlier photographic plate focal-plane detectors (**Fig. 9**).

A further important utilization of the nuclear triodes has followed their fabrication in an annular geometry similar to that shown in Fig. 7. With radial position sensitivity it becomes possible to correct on-line, and event by event, for the kinematic variation of particle energy with angle over the aperture of the detector. Without this correction possibility all

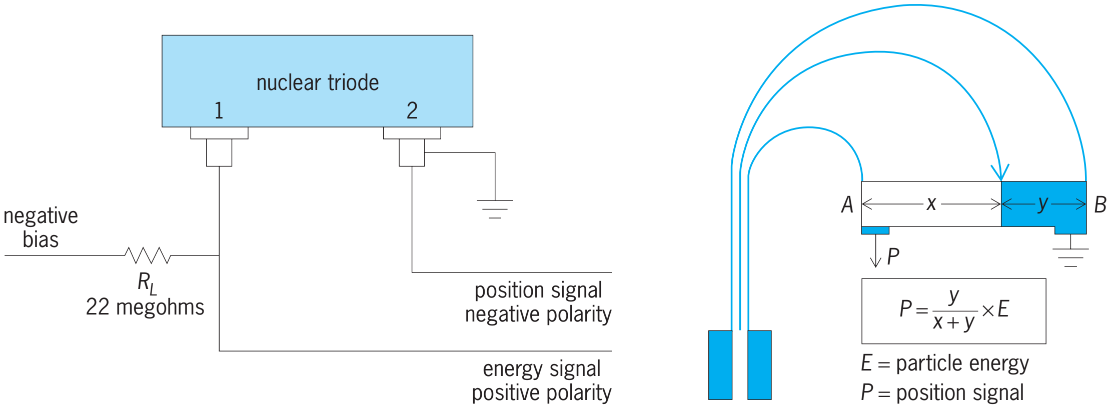


Fig. 9. Schematic of position-sensitive (nuclear triode) detector in focal plane of 180° magnetic spectrograph.

particle group structures in the detector spectrum are smeared beyond recognition. *See* PARTICLE DETECTOR; SEMICONDUCTOR. James M. McKenzie

Bibliography. D. A. Bromley, Detectors in nuclear science, *Nucl. Instrum. Meth.*, 162:1–8, 1979; G. T. Ewan, The solid ionization chamber, *Nucl. Instrum. Meth.*, 162:75–92, 1979; E. Laegsgaard, Position-sensitive semiconductor detectors, *Nucl. Instrum. Meth.*, 162:93–111, 1979; J. M. McKenzie, Development of the semiconductor radiation detector, *Nucl. Instrum. Meth.*, 162:49–73, 1979.

Junction diode

A semiconductor rectifying device in which the barrier between two regions of opposite conductivity type produces the rectification (**Fig. 1**). Junction diodes are used in computers, radio and television, brushless generators, battery chargers, and electrochemical processes requiring high direct current and low voltage. Lower-power units are usually called semiconductor diodes, and the higher-power units are usually called semiconductor rectifiers. For a discussion of conductivity types, carriers, and impurities *see* SEMICONDUCTOR.

Junction diodes are classified by the method of preparation of the junction, the semiconductor material, and the general category of use of the finished device. By far the great majority of modern junction diodes use silicon as the basic semiconductor material. Germanium material was used in the first decade of semiconductor diode technology, but has given way to the all-pervasive silicon technology, which allows wider temperature limits of operation and produces stable characteristics more easily. Other materials are the group 13–15 compounds, the most common being gallium arsenide, which is used where its relatively large band-gap energy is needed. A partial list of silicon types includes the diffused silicon switching diode, alloyed silicon voltage reference diode, epitaxial planar silicon photodiode, and diffused silicon rectifier. Other types include the ion-implanted varactor diode and the gallium arsenide light-emitting diode.

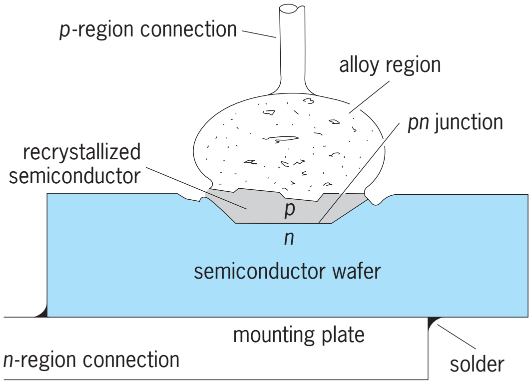


Fig. 1. Section of a bonded or fused junction diode.

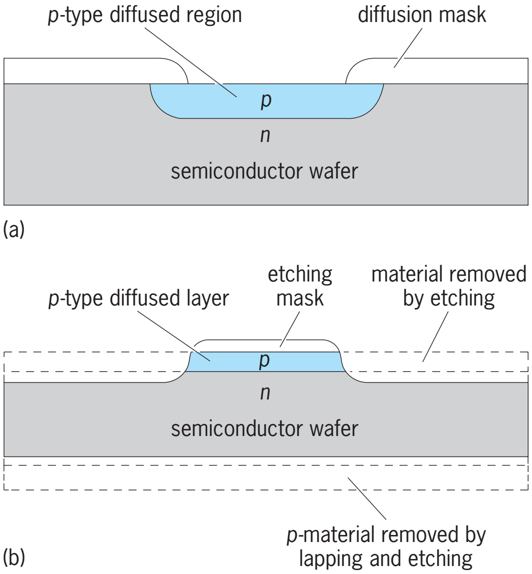


Fig. 2. High-speed diffused silicon diodes. (*a*) Mesaless structure. (*b*) Mesa structure.

In silicon units nearly all categories of diodes are made by self-masked diffusion, as shown in **Fig. 2*a***. Exceptions are diodes where special control of the doping profile is necessary. In such cases, a variety of doping techniques may be used, including ion implantation, alloying with variable recrystallization rate, silicon transmutation by neutron absorption, and variable-impurity epitaxial growth. The mesa structure shown in Fig. 2*b* is used for some varactor and switching diodes if close control of capacitance and voltage breakdown is required. *See* ELECTRONIC SWITCH; RECTIFIER; SEMICONDUCTOR DIODE.

Fabrication methods. The alloy and mesa techniques are largely historical, but were important in the development of junction diodes. The alloy junction section (Fig. 1) is produced by placing a pill of doped alloying material on the clean flat surface of a properly oriented semiconductor wafer and heating it until the molten alloy dissolves a portion of the semiconductor immediately beneath it. Upon cooling, the dissolved semiconductor, now containing the doping impurity, recrystallizes upon the surface of the undissolved semiconductor, reproducing its crystal structure and creating a *pn* junction at the position marking the limit of the solution of the original wafer. If such a junction is held at the peak temperature of its alloying cycle for sufficient time to allow diffusion of the alloy impurity beyond the limit of the dissolved semiconductor into the solid semiconductor, the junction produced is called alloy-diffused.

The planar diffused junction section (Fig. 2*a*) is produced in silicon by first polishing the top surface of a large silicon wafer and then oxidizing the surface by heating the wafer at about 1000°C (1800°F) in the presence of wet oxygen. After about 0.5 micrometer of oxide has grown on the surface, the wafer is cooled, and an array of holes is opened in the oxide by high-precision etching geometrically controlled by a photoresist technique. A very heavily doped thin oxide layer is chemically deposited in the holes opened in the oxide. This predeposition step is followed by a drive-in diffusion at a higher temperature, causing the deposited impurity to penetrate the substrate, thereby forming diffused *pn* junctions beneath each hole. Subsequently the individual junctions are separated out of the large wafer by scribing and breaking and are encapsulated as individual diodes. Such planar diffused diodes have relatively high breakdown voltages and low leakage currents. The ends of the junction are automatically protected by the oxide mask so that such diodes show long-term stability. This protection by the oxide is often referred to as passivation.

Planar diodes and planar transistors are used in integrated circuits. The diodes in integrated circuits usually consist of the emitter junction or collector junction of a transistor structure rather than being fabricated as a separate diode. Most discrete diodes are power rectifiers, voltage regulators, varactors, or light-emitting diodes. *See* INTEGRATED CIRCUITS; VARACTOR; VOLTAGE REGULATOR.

The mesa structure (Fig. 2*b*) is produced by diffusing the entire surface of the large wafer and then delineating the individual diode areas by a photoresist-controlled etch that removes the entire diffused area except the island or mesa at each diode site.

Still another method of doping control used in modern diodes is through epitaxially deposited material. In this process the polished wafer is subjected at an elevated temperature to a vapor containing a compound of the semiconductor together with a compound containing the appropriate doping element. These compounds decompose upon contact with the surface of the wafer and cause the semiconductor to grow a layer of doped material on its surface. Under proper conditions of cleanliness and growth rate, the underlying crystal structure is propagated into the growing layer, which is then said to be epitaxial in character. In this way either localized areas or entire surfaces of either conductivity type may be produced. In diode fabrication it is typical to use the epitaxially grown material as a lightly doped layer over the original substrate material of the same conductivity type. The junction is then formed in the lightly doped layer by masked diffusion of the opposite-conductivity-type material. By this means the thickness of the web of lightly doped material immediately beneath the diffusion can be controlled to give both a desired reverse breakdown voltage and a relatively constant capacitance. Forward-bias recovery time can be controlled in a trade-off with reverse breakdown voltage in such a structure.

A method of doping control used when special doping concentration profiles are needed, or when localized doping must be accomplished without self-masking oxide, is ion implantation. At present the largest use of this technique in *pn* junction fabrication is to replace the chemical predeposition step in the planar diffusion process. Here ion implantation gives a much more precise control of the sheet resistivity of the diffusion, and it can be accomplished without opening holes in the protective oxide. Crystal damage is automatically healed during the subsequent drive-in diffusion. *See* ION IMPLANTATION.

Junction rectification. Rectification occurs in a semiconductor wherever there is a relatively abrupt change of conductivity type. In any semiconductor the product of the concentrations of the majority and minority current carriers is a temperature-dependent equilibrium constant. The conductivity is proportional to the majority carrier concentration and inversely proportional to the minority-carrier concentration. When a *pn* junction is reverse-biased (*p*-region negative with respect to the *n*-region), the majority carriers are blocked completely by the barrier, and only the minority carriers can flow under the barrier. This minority carrier current is the sum of the individual currents from the *n*- and *p*-regions, and each component is inversely proportional to the conductivity of its region. In addition, there is a thermal regeneration current of minority carriers generated in the depletion region of the reverse-biased junction. In silicon the regeneration current dominates and is about 10^{-3} A/m^2 at room temperature.

When a *pn* junction is forward-biased (*p*-region positive with respect to the *n*-region), the majority hole and electron distributions can flow into the opposite region because the bias has markedly lowered the barrier. Since electrons flowing into a *p*-region or holes flowing into an *n*-region represent a great increase in minority-carrier concentration, the thermodynamic equilibrium of the holes and electrons is disturbed, and the product of their concentrations increases as the junction is approached. The resistivity of both the *n*- and *p*-type regions is considerably lowered by these excess minority carriers, and the forward current is greater than the current through a geometrically equivalent bar of material containing no *pn* junction.

The electrons in an *n*-type semiconductor are given up to the conduction process by donor impurity atoms which remain as fixed, positively charged centers. Similarly, the holes of a *p*-region are created by the capture of electrons by acceptor impurity atoms which remain as fixed, negatively charged centers. In both cases the space charge of the ionized impurity centers is neutralized by the space charge of the majority carriers.

At a *pn* junction the barrier that keeps majority carriers away consists of a dipole layer of charged impurity centers, positive on the *n*-type side and negative on the *p*-type side. When a reverse bias is applied, the barrier height increases and requires more charge in the dipole layer to produce the required step in voltage. To add to the charge, the layer must widen, because ionized impurities are in fixed positions in the crystal. As the layer widens, the capacitance of the junction decreases since the plates of the capacitor are farther apart. Therefore, a *pn* junction acts as a variable capacitance as well as a variable resistance. In this application, it is called a varicap or a varactor diode.

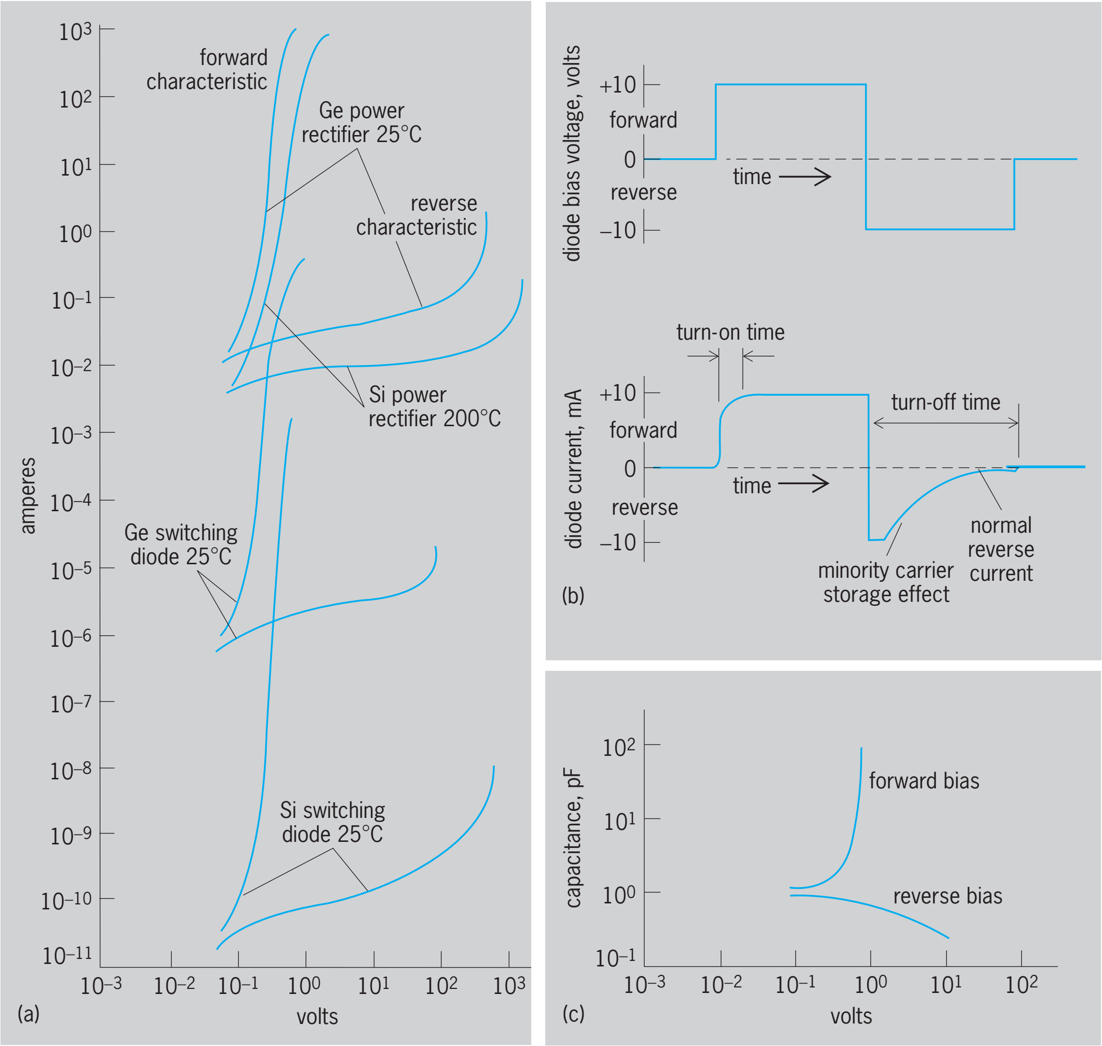


Fig. 3. **Junction diode characteristics. (*a*) Rectification. (*b*) Switching. (c) Silicon switching diode capacitance. °F = (°C × 1.8) + 32.**

Optical properties. When light of sufficient energy is absorbed by a semiconductor, excess minority carriers are created. In a *pn* junction device these excess carriers will increase the reverse-bias leakage current by a large factor if they are within diffusion distance of the junction. If the junction is open-circuited, a forward voltage will develop to oppose the diffusion of the excess carriers generated by the light absorption. This photovoltaic response is the basis of the operation of solar cells and most photographic exposure meters. *See* EXPOSURE METER; PHOTOVOLTAIC CELL; PHOTOVOLTAIC EFFECT; SOLAR CELL.

The inverse of the above effect also exists. When a *pn* junction in a so-called direct-gap semiconductor is forward-biased, the electrically injected excess minority carriers recombine to generate light. This is the basis of light-emitting diodes and injection lasers. Typical direct-gap semiconductors (suitable for light-emitting diodes) are compounds between 13 and 15 group elements of the periodic table such as gallium arsenide. *See* LASER.

Characteristics. Typical rectification, switching, and capacitance characteristics of a junction diode are shown in **Fig. 3**. Rectification characteristics (Fig. 3*a*) show that silicon units provide much lower reverse leakage currents and higher voltage breakdowns and can operate up to 200°C (392°F).

For switching purposes, turn-on and turn-off times are most important (Fig. 3*b*). The turn-on time of a diode is governed by its junction capacitance and is usually short. The turn-off time, usually the critical characteristic, is governed by the time required to remove all of the excess minority carriers injected into the *n*- and *p*-regions while the diode was in the forward-bias state. This is called the minority carrier storage effect, and it is of the order of a few microseconds for good switching diodes. Silicon diodes are usually somewhat superior to germanium units in this respect. The limits of operation of present junction diodes are about 2500 V reverse-standoff voltage and 1500 A forward current in power rectifiers; about 1.0 nanosecond reverse recovery time and 100 picoseconds rise time for fast-switching diodes; a minimum reverse leakage current in a small signal diode is about 0.01 nA.

For further discussion of the properties of *pn* junctions *see* JUNCTION TRANSISTOR; TRANSISTOR.

Lloyd P. Hunter

Bibliography. J. Singh, *Semiconductor Devices: An Introduction*, 1994; S. M. Sze, *Semiconductor Devices*, 1985; E. S. Yang, *Microelectronic Devices*, 1988; M. Zambuto, *Semiconductor Devices*, 1989.

Junction transistor

A transistor in which emitter and collector barriers are formed by *pn* junctions between semiconductor regions of opposite conductivity type. These junctions are separated by a distance considerably less than a minority-carrier diffusion length, so that minority carriers injected at the emitter junction will not recombine before reaching the collector barrier and therefore be effective in modulating the collector-barrier impedance. Junction transistors are widely used both as discrete devices and in integrated circuits. The discrete devices are found in the high-power and high-frequency applications. Junction transistors range in power rating from a few milliwatts to about 300 W, in characteristic frequency from 0.5 to 2000 MHz, and in gain from 10 to 50 dB. Silicon is the most widely used semiconductor material, although germanium is still used for some applications. Junction transistors are applicable to any electronic amplification, detection, or switching problem not requiring operation above 200°C (392°F), 700 V, or 2000 MHz. Not all these limits can be achieved in one device, however. Junction transistors are classified by the number and order of their regions of different conductivity type, by the method of fabricating and structure, and sometimes by the principle of operation. Most modern transistors are fabricated by the silicon self-masked planar double-diffusion technique. The alloy technique and the grown-junction technique are primarily of historical importance. For a general description and definition of terms used here and a description of the mechanism of operation *see* TRANSISTOR.

Alloy-junction transistors. Also called fused-junction transistors, these are made in the *pnp* and *npn* forms. The emitter and collector regions are formed by recrystallization of semiconductor material from a solution of semiconductor material dissolved in some suitably doped metal alloy. The major metal of the alloy serves as the solvent for the semiconductor, while the minor element serves as a source for doping impurity in order to render the recrystallized material opposite in conductivity type to the original wafer.

Alloy junctions are abrupt and allow for bidirectional operation. They usually show a low series resistance, and were therefore used in high-power transistors.

Figure 1 compares several transistor profiles which show how the impurity content varies through the structure. In these profiles C_p is the concentration of the *p*-type impurity; C_n is the concentration of the *n*-type impurity. The net impurity content determines the conductivity type and magnitude. The profile of the alloy transistor shows that there are abrupt changes of impurity concentration at emitter and collector junctions and that the conductivities of emitter and collector regions are therefore high compared to those of the base region. Such a structure shows good emitter-injection efficiency but only moderate collector-voltage rating and relatively high collector capacitance. *See* SEMICONDUCTOR.

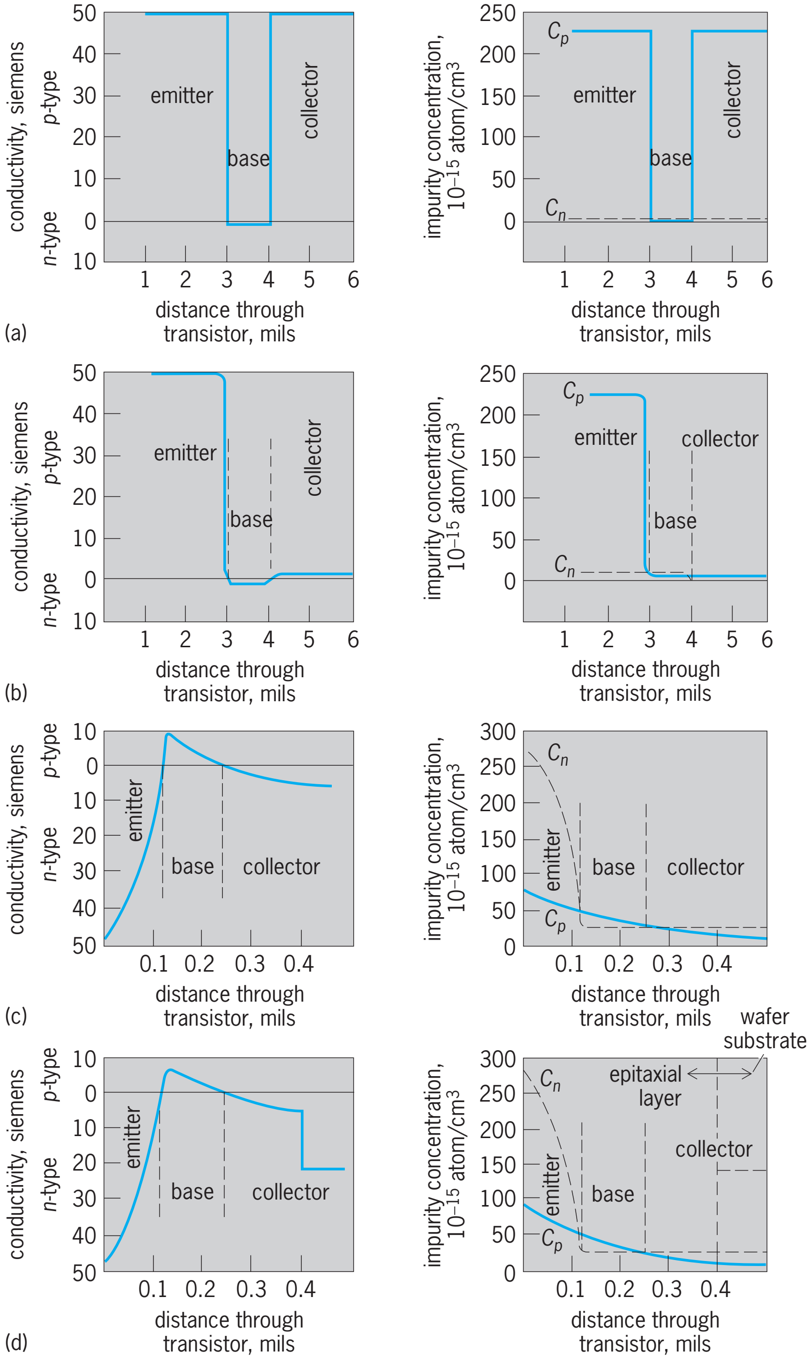


Fig. 1. Conductivity and impurity profiles of typical junction transistors. (*a*) *pnp* alloy-junction type. (*b*) *pnp* grown-junction type. (*c*) *npn* double-diffused-junction type. (*d*) *npn* epitaxial double-diffused-junction type. 1 mil = 25.4 μm.

Grown-junction transistors. These are made in the *pnp* and *npn* forms, as well as in more complicated forms. There are several variations of the grown-junction technique. The simplest consists of successively adding different types of impurities to the melt from which the semiconductor crystal is being grown.

A semiconductor crystal is usually grown by dipping the end of a seed crystal into molten semiconductor and by arranging the thermal gradients so that new semiconductor solidifies on the end of the seed as it is slowly withdrawn. The solid-liquid interface is roughly a plane perpendicular to the axis of withdrawal. A *pnp* structure can be grown by starting with a *p*-type melt; by adding, at one point in the crystal growth, enough *n*-type impurity to give a slight excess over the *p*-type impurity originally present; and, after growth has continued for a few micrometers, by adding an excess of *p*-type impurity. The last-grown region will be the emitter region, and the original *p*-type crystal will be the collector region. The impurity profile of such a structure is shown in Fig. 1*b*.

The high-conductivity emitter region gives a good injection efficiency, and the junction between the base and collector regions is gradual enough so that the unit will show a relatively low collector capacitance and a high breakdown voltage. The one disadvantage of this method is that both the collector and base regions show relatively high series resistances.

Planar diffused epitaxial transistors. The structure of this transistor is shown in section in **Fig. 2**, and the doping profile through the emitter, base, and collector is shown in Fig. 1*d*. In this structure both collector and emitter junctions are formed by diffusion of impurities from the top surface, as shown in Fig. 2. Using silicon, the structure is formed by growing a diffusion mask of native oxide (silicon dioxide) on the previously polished wafer. A hole is opened in the oxide by a photoresist etch technique (Fig. 2*a*) to define the area of the collector buried layer. For a *p*-type substrate a heavy concentration (n^+) of *n*-type impurity such as phosphorus is diffused into the substrate through the opening in the masking oxide. The oxide is etched away, and an epitaxial layer of lightly doped (n^-) silicon is grown over the entire wafer by vapor decomposition at a temperature low enough to prevent significant diffusion of the n^+ material out of the buried layer (Fig. 2*b*). A new oxide layer is grown on the surface of the epitaxial layer, and an opening is etched in it to define the limits of the *p*-type base diffusion (Fig. 2*c*). (This automatically controls the collector junction geometry and capacitance.) The masking oxide is again stripped and regrown for the masking of the n^+ diffusion used to form the emitter and collector contact region (Fig. 2*d*). Next the emitter mask is removed, and an impervious layer of oxide is formed over the surface of the crystal. A layer of glass is bonded to the crystal by means of the oxide layer. The glass must match the expansion coefficient of silicon fairly well, and the oxide must be sufficiently impervious to the glass at the bonding temperature to prevent the diffusion of impurities from the glass into the silicon transistor structure. Finally, holes are etched in the glass-oxide structure so that electrical contact can be made to the various regions of the transistor (Fig. 2*e*).

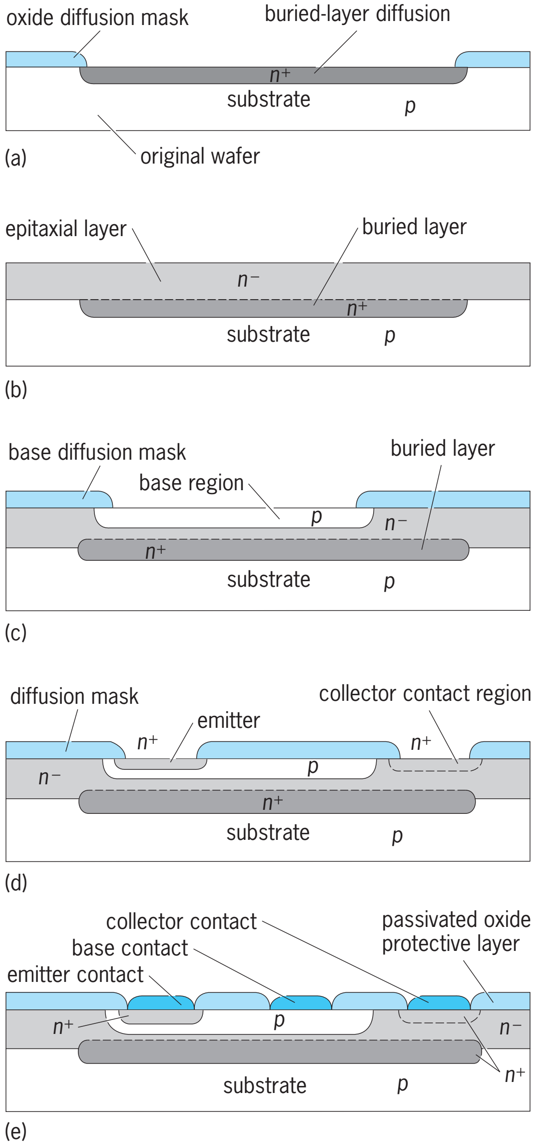


Fig. 2. Double-diffused planar epitaxial transistor structure and method of fabrication. (*a*) Buried layer. (*b*) Epitaxial layer. (*c*) Collector junction formation. (*d*) Emitter junction. (*e*) Contact stripe placement.

In modern technology the above-described base and emitter diffusions are carried out in two steps: a predeposition step, in which a very thin layer of heavily doped oxide is chemically deposited over the open surface of the silicon in the hole opened in the masking oxide; and a drive-in diffusion step, in which the deposited dopant is diffused into the silicon at a higher temperature than that used for the predeposition. This typically controls the sheet resistance of the final diffusions to about ±10% of the design value. The forward current-transfer ratio of the transistor is determined by the ratio of these sheet resistances through the medium of the injection efficiency of minority carriers.

The chemical predeposition step is being replaced by ion implantation directly through the oxide. The masking is accomplished by placing a layer of photoresist on top of the oxide. This eliminates the oxide etching step and allows an accurately metered deposition by controlling the time of bombardment and the current of the ion beam. This modification of the process promises to keep the emitter and base region sheet resistivities within about ±1% of the design value. *See* ION IMPLANTATION.

In this transistor, formation of the base region by diffusion from the emitter side produces a steep doping gradient and thereby a strong electric field in the base region. In the typical alloy-junction transistor (uniform base doping) the minority-carrier transport across the base is achieved by a relatively slow diffusion process. In this diffused base type (sometimes called a drift transistor) the base region shows a high conductivity gradient, decreasing from the emitter to the collector (Fig. 1*d*). This conductivity gradient means that the majority-carrier concentration is much greater near the emitter than near the collector. In order to cancel the natural diffusion of majority carriers from the high- to the low-concentration region, an electric field must exist of such a polarity as to tend to drive majority carriers back toward the emitter. This polarity of field then tends to drive minority carriers from the emitter to the collector; when normal bias is applied to the device, excess injected minority carriers will be accelerated across the base by this field.

The buried layer of n^+-doped material has very low resistance and acts as a shorting bar between the area immediately beneath the emitter and the area immediately beneath the collector contact stripe, thus maintaining a low collector series resistance even if the n^- material of the epitaxial region is of quite high resistivity. The collector breakdown voltage may be maintained at a reasonably high value and the collector capacitance at a low value by controlling the thickness of the n^- material between the base and the buried layer and by keeping the doping level of the n^- material quite low.

Mesa transistors. These transistors minimize the collector capacitance by limiting the collector junction area. This area limitation is achieved by etching away the surrounding material so that the entire transistor structure stands up from the body of the wafer like a small mesa. This structure gave the highest frequency response for many years. It is now replaced by the planar type of structure.

Power transistors. These are used in the output stage of an electronic circuit both as switches and as amplifiers. Depending on the load, a high voltage rating, a high current rating, or a high power rating may be required. With any of these, heat dissipation within the device is a serious limitation. In order to obtain high current capability in a power transistor, a large emitter junction area is required. The base-region recombination current produces a lateral voltage drop between the center of the emitter area and the center of the base contact area in a planar device. This voltage tends to bias the center of the emitter area to off and concentrate the injection current at the periphery of the emitter. Modern silicon power transitors minimize this effect by using emitter junctions with a large perimeter-to-area ratio, usually in the form of a multipronged fork with base contacts interdigitated between the tines. This preserves high forward current-transfer ratio to large values of emitter current. *See* CONTROLLED RECTIFIER.

Unijunction transistor. This device is really a diode in which the resistance of a portion of the base region is modulated by minority carriers injected by forward-biasing its single junction. Its structure typically consists of a lightly doped base region with ohmic contacts at opposite ends. The single junction is formed over a narrow range near the center of the base region by a shallow diffusion of heavily doped material of the opposite conductivity type. If a bias current is set up from end to end in the base, the potential at the junction can be set at a desired reverse bias relative to ground. If a signal is applied to the junction electrode, the device will turn on when the signal exceeds the original reverse-bias potential of the base at that point. Once forward-biased, the junction injects sufficient minority carriers into the base to short the region beneath the junction to the ground end of the base, and the device remains conducting until reset, either by the base bias or by the emitter signal. These devices show a typical negative resistance characteristic and are used for timing, control, and sensing circuits.

Summary. Silicon planar passivated transistors show a wide range of performance with characteristic frequencies up to 2000 MHz, voltage ratings of 12–700 V, and power dissipation ratings of 100 mW–300 W. The highest-frequency devices range up to 4000 MHz. Silicon planar technology is used in fabricating integrated circuit chips. The general form of the transistor structure displayed in Fig. 2 is used in integrated circuits. Such a structure is used for diodes as well as transistors since, for example, it is necessary only to connect the base and collector contacts to use the collector junction as a diode. *See* INTEGRATED CIRCUITS.

Lloyd P Hunter

Bibliography. J. J. Brophy, *Basic Electronics for Scientists*, 5th ed., 1990; G. W. Neudeck, *The Bipolar Junction Transistor*, 2d ed., 1989; S. M. Sze, *Semiconductor Devices*, 1985; E. S. Yang, *Microelectronic Devices*, 1988.

Jungermanniales

The largest order of liverworts, often called the leafy liverworts; it consists of 43 families. The leaves are in three rows, with the underleaves usually reduced or lacking. Other distinctive features include a perianth formed by a fusion of modified leaves, a short-lived seta, and a four-valved capsule. The leaves have an embryonic bilobed phase which may be lost on further development.

The plants of this order are dorsiventrally organized and leafy. They grow by means of an apical cell with three cutting faces, resulting in two rows of lateral leaves and a third row of underleaves which are generally reduced, and sometimes lacking. The stems lack a central strand. A medulla of delicate cells is generally surrounded by a cortex. Rhizoids are usually present, all smooth. The leaves pass through a primordial two-lobed stage but may become two- to several-lobed, or unlobed (owing to obliteration of one primordial lobe). A midrib is lacking. The cells often have corner thickenings, and oil bodies are usually present in all green cells. Asexual reproduction by gemmae is common. Antheridia occur in leaf axils, sometimes also in axils of underleaves. The stalk is usually 1-2- or less commonly 4-6-seriate. Archegonia are terminal and have a 5-seriate neck. The sporophyte is usually protected by a perianth (in addition to a calyptra) formed by the fusion of leaves. The seta, usually long, consists of delicate, hyaline cells. The capsule is four-valved, with a wall of 2–10 layers of cells. The elaters have spiral-thickened walls. The haploid chromosome number is 8 or 9. *See* BRYOPHYTA; HEPATICOPSIDA; JUNGERMANNIIDAE.

Howard Crum

Bibliography. M. Fulford, Manual of the leafy Hepaticae of Latin America, *Mem. N.Y. Bot. Gard.*, 11:1–172 (1963), 173–276 (1966), 277–392 (1968), 393–535 (1976); R. M. Schuster, *The Hepaticae and Anthocerotae of North America East of the Hundredth Meridian*, vols. 1–4, 1966–1980.

Jungermanniidae

One of the two subclasses of liverworts (class Hepaticopsida). The plants may be thallose, with little or no tissue differentiation, or they may be organized into erect or prostrate stems with leafy appendages. The leaves, generally one cell in thickness, are mostly arranged in three rows, with the third row of underleaves commonly reduced or even lacking. Oil bodies are usually present in all of the cells. The rhizoids are smooth. The capsules, generally dehiscing by four valves, are usually elevated on a long, delicate, short-lived seta. The spore mother cells are deeply lobed.

The subclass consists of the orders Takakiales, Calobryales and Jungermanniales, which are leafy, and the Metzgeriales, which are mostly thallose. *See* BRYOPHYTA; CALOBRYALES; HEPATICOPSIDA; JUNGERMANNIALES; METZGERIALES; TAKAKIALES.

Howard Crum

Bibliography. H. Inoue, *Illustrations of Japanese Hepaticae*, vol. 1, 1974; R. M. Schuster, Phylogenetic and taxonomic studies on Jungermanniidae, *J. Hattori Bot. Lab.*, 36:321–405, 1972.

Jupiter

The largest planet in the solar system, and the fifth in the order of distance from the Sun. It is visible to the unaided eye except for short periods when in near conjunction with the Sun. Usually it is the second brightest planet in the sky; only Mars at its maximum luminosity and Venus appear brighter.

Planet and its orbit. The main orbital elements are a semimajor axis, or mean distance to the Sun, of 484×10^6 mi (778×10^6 km); an eccentricity of 0.0489, causing the distance to the Sun to vary about 47×10^6 mi (75×10^6 km) between perihelion and aphelion; sidereal revolution period of 11.86 years; mean orbital velocity of 8.1 mi/s (13.1 km/s); and inclination of orbital plane to ecliptic of 1.3°. *See* PLANET.

The apparent equatorial diameter of its disk varies from about 47″ at mean opposition to 32″ at conjunction. The polar flattening due to its rapid rotation is considerable and is easily detected by visual inspection; the ellipticity is $(r_e - r_p)/r_e = 0.065$, where r_e is the equatorial radius and r_p is the polar radius. The equatorial diameter is about 88,850 mi (142,984 km), and the polar diameter is 88,086 mi (133,708 km). The volume is about 1321 (Earth = 1), and the mass is about 317.5 (Earth = 1). The mean density is 1.33 g/cm^3, a low value characteristic of the four giant planets. The mean acceleration of gravity at the visible surface is about 85 ft/s^2 (26 m/s^2); however, the centrifugal force at the equator reduces the effective acceleration of gravity to about 78 ft/s^2 (24 m/s^2).

Phases. As an exterior planet, Jupiter shows only gibbous phases and full phase from Earth. Because of the large size of Jupiter's orbit compared with that of the Earth, the maximum phase angle is only 12° at quadratures, and the phase effect shows up only as a slightly increased darkening of the edge at the terminator. The apparent visual magnitude at mean opposition is −2.4, and the corresponding value of the reflectivity (geometrical albedo) is about 0.3; the physical albedo is 0.34. The high value of the albedo, characteristic of the four giant planets, indicates the presence of a dense, cloud-laden atmosphere. *See* ALBEDO.

Telescopic appearance. Through an optical telescope Jupiter appears as an elliptical disk, strongly darkened near the limb and crossed by a series of bands parallel to the equator (**Fig. 1**). Even fairly small telescopes show a great deal of complex structure in the bands and disclose the rapid rotation of the planet. The period of rotation is about 9 h 55 m, the shortest of any planet. The features observed, however, do not correspond to the solid body of a planet but to clouds in its atmosphere, and the rotation period varies markedly with latitude. The rotation period of any given zone is not exactly constant but suffers continual fluctuations about a mean value. Occasionally, short-lived atmospheric phenomena may depart more strongly from the mean rotation period of the zone in which they appear, and thus drift rapidly with respect to other details in the zone. The rotation axis is inclined only 3° to the perpendicular to the orbital plane, so that seasonal effects are practically negligible.

Red Spot. Apart from the constantly changing details of the belts, some permanent or semipermanent markings have been observed to last for decades or even centuries, with some fluctuations in visibility. The most conspicuous and permanent marking is the

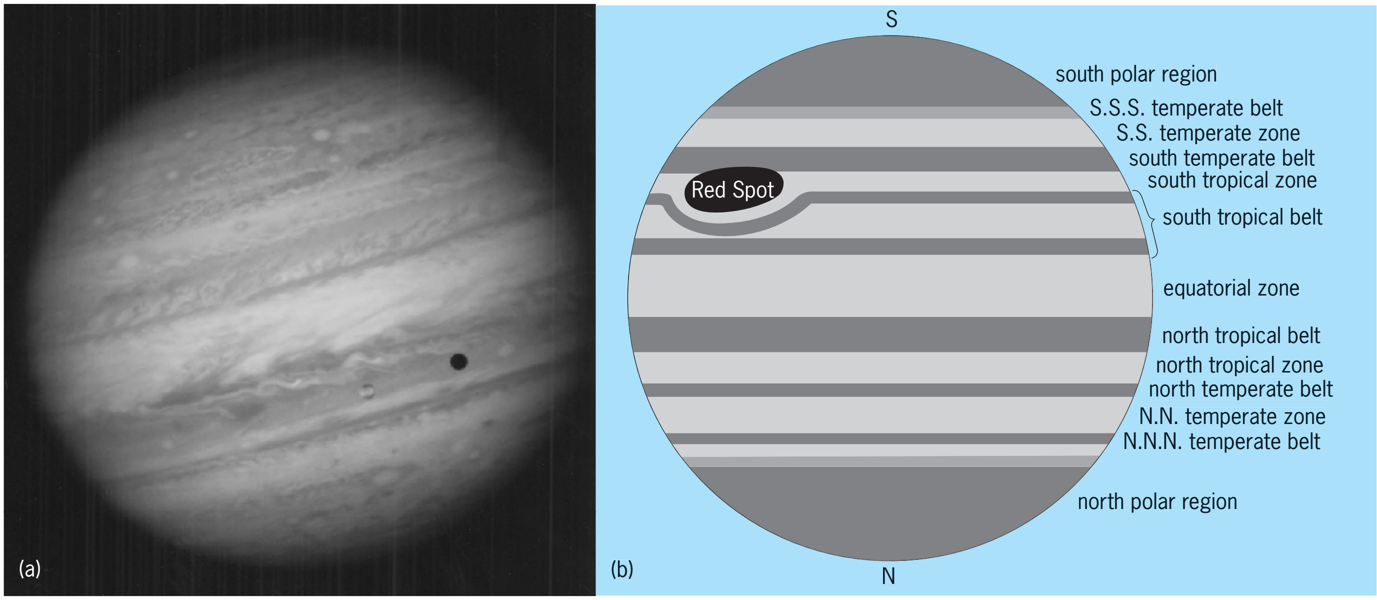


Fig. 1. Jupiter. (*a*) Telescopic appearance from the *Hubble Space Telescope* (*Space Telescope Science Institute: Jet Propulsion Laboratory; NASA*). (*b*) Principal bands.

great Red Spot, discovered by J. D. Cassini in 1665 and observed continually since 1878, when it gained its present red color. At times it has been conspicuous and strongly colored; at other times it has been faint and only slightly colored, and occasionally only its outline or that of the bright "hollow" of the south temperate zone which surrounds it has remained visible.

The current dimensions of the Red Spot are 15,500 mi (25,000 km) in longitude and 7500 mi (12,000 km) in latitude, though the longitudinal dimension varies from 25,000 mi (40,000 km) to 9000 mi (14,000 km). The Spot has been shrinking since 1943 at a rate of 0.19° of longitude per year. It also drifts in longitude by 4.5° per year. It has an essentially calm center, surrounded by a restraining "collar" of high-speed winds which rotate counterclockwise in 6–10 days. This anticyclonic storm (rotation in opposite direction to the planet's) consists of an elevated central region of ammonia gas and ice clouds (**Fig. 2**). Wind speed increases with depth. The Spot is surrounded by towering water-vapor-based thunderstorms, with the deepest clouds in the surrounding collar. The total depth is probably 12–25 mi (20–40 km). These clouds are about 18°F (10°C) colder than the surrounding weather. The Spot's distinctive coloration is probably due to chemical compounds (perhaps containing phosphorus) transported from deep within the atmosphere.

The origin and longevity of the Great Red Spot remain difficult to explain. The vortex is probably not attached to any solid surface feature below the clouds. It is thought to be an eddy of atmospheric gases driven by the strong Coriolis force that stems from the planetary rotation. Atmospheric perturbations or whirlpools coalesce as they pass in the underlying jet stream, forming a vortex of increasing size, which can maintain stability for some time. Several smaller and similarly colored versions of the Spot have been observed, but none has attained the size or lifetime of the original. *Galileo* probe observations suggest that the energy necessary to sustain Red Spot survival comes from the Spot's regular consumption of these eddies several thousand kilometers in diameter, which carry up convective energy from below. This implies a possible lifetime around 100,000 year. *See* CORIOLIS ACCELERATION.

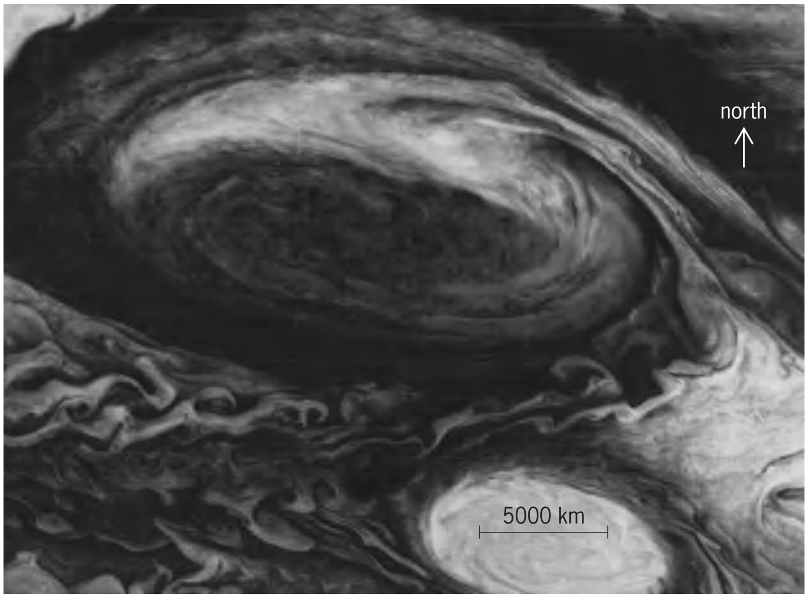


Fig. 2. Details of the Red Spot as seen by the *Voyager 1* fly by. Features include a white oval with a wake of counterrotating vortices, puffy features inside the Red Spot, and reverse-S spirals inside both the Red Spot and the oval. The large white feature extending over the northern part of the Red Spot was observed to revolve about the Red Spot center with a period of 6 days. 5000 km = 3000 mi. (*NASA*)

Atmosphere. Jupiter's visible belts and zones reflect the complicated vertical atmospheric structure. The atmosphere consists primarily of hydrogen (H_2), helium (He), ammonia (NH_3), methane (CH_4), water (H_2O), hydrogen sulfide (H_2S), and other trace compounds such as acetylene ($HC{\equiv}CH$), ethane (CH_3CH_3), hydrogen cyanide (HCN), hydrogen deuteride (HD), carbon monoxide (CO), and phosphine (H_3P). The highest atmospheric features are ammonia crystal haze layers, variable in appearance and density. The visible cloud tops are formed of aerosols of mostly ammonia, with little water vapor, akin to cirrus clouds in Earth's atmosphere. As depth increases, cloud particles become ammonia crystals and ammonium hydrosulfide (NH_4SH) droplets, followed by cumulus clouds of water ice crystals and droplets (**Fig. 3**).

All the compounds previously mentioned are colorless; therefore, the strong yellow, red, blue, and brown colorations seen in the clouds must be due to other agents. Perhaps phosphorus and sulfur compounds create the yellows and reds, while the blue features may be hydrogen made visible by Rayleigh scattering. Infrared observations have found the highest features to be red, followed by white, brown, and blue.

The *Galileo* probe found unexpected results in its plunge into the upper Jovian atmosphere. No distinct water cloud was observed, and less than 10% of the expected abundance of water vapor was seen, although two distinct ammonia and ammonium hydrosulfide cloud levels were seen. It is thought that the probe may have entered an unusually clear and dry area of the atmosphere, a figurative desert. More recent observations give an overall water content similar to solar values. When signal was lost, the probe had penetrated 0.22% of the Jovian atmosphere (about 43 mi or 70 km) where the temperature was 1970°F (1350 K, hotter than expected). This excess heat may have been caused by gravity waves from the lower atmosphere or else heating by soft electron collisions on the hydrogen (H_2) molecules. The ammonia cloud was very tenuous and inhomogeneous near the probe entry point. An ammonia-sulfur cloud was encountered well below its expected height, pointing out the extreme variability of the atmosphere. The probe also found helium raining from upper to lower depths. The *Galileo* probe discovered anomalously high quantities of argon, krypton, and xenon, suggesting that Jupiter had to trap them physically by condensation or freezing. This would require Jovian formation farther out in the solar system than its present location, possibly in the Kuiper Belt. *See* KUIPER BELT.

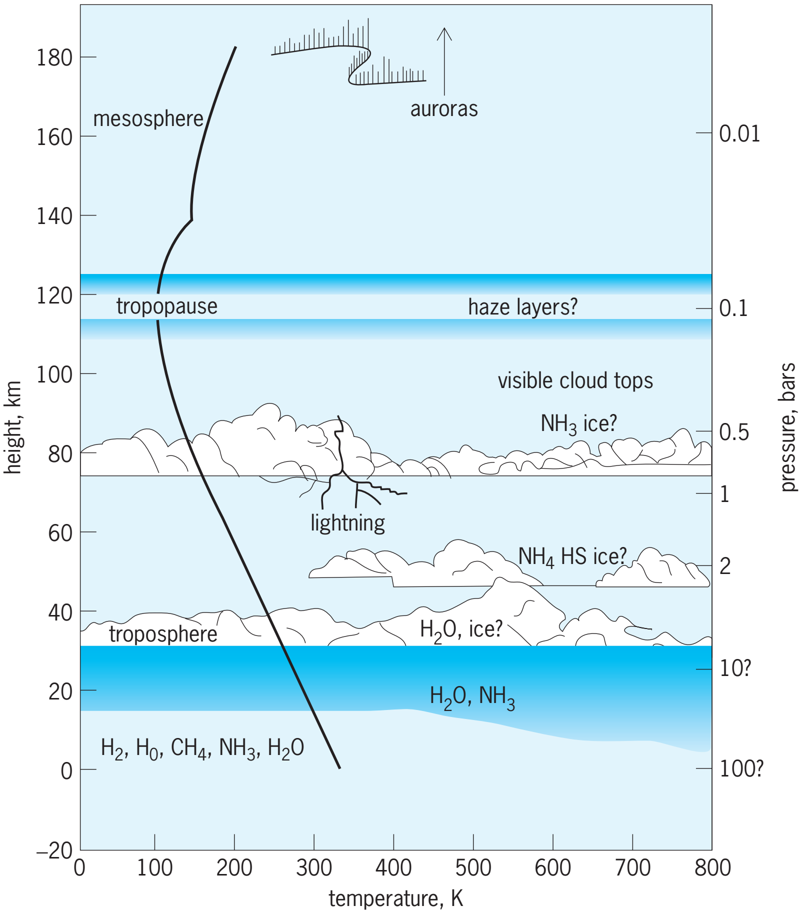


Fig. 3. Diagram showing the locations of the cloud decks of Jupiter. 1 km = 0.6 mi.

The atmospheric structure displays latitudinal variations. Clouds near the equator are lifting and spreading toward higer latitudes, as in the horse latitudes on Earth. Atmospheric aerosols subside at the equator and upwell at high latitudes. The mean temperature of the Jovian disk is −259°F (129 K), but bands of high temperature are found at the poles and mid-latitudes. The equatorial regions alternate between warm and cold over a cycle of 4–5 years.

Zones and belts. The belts and zones represent regions of differing cloud altitudes and compositions. The atmosphere is divided by a series of prograding and retrograding jet streams. Convection is strong since the bright zones are cooler and higher by 9–12 mi (15–20 km) than the dark belts, with their higher albedos arising from solid ammonia crystals. The zones are regions of upward motion and high pressure. *Galileo* found wind speeds of 0.1–0.15 mi/s (160–220 m/s) to extend to unexpected depths, as well as supersonic turbulence in the upper atmosphere on the order of 0.6–6 mi/s (1–10 km/s). This is similar to motions in stellar, not planetary, atmospheres, suggesting particularly stormy upper regions. The energy to maintain such high-velocity winds likely comes from fluid-mechanical convective processes from the interior heat source (discussed below), with some contribution from lightning. The equatorial zonal jet streams have not sensibly changed in over 80 years of observation, despite vigorous turbulence in the defining shear zones.

Belts are aerosol-free and dry; zones are cloudy and ammonia-rich. The motions in the belts are cyclonic, and anticyclonic in the zones. The anticyclones have smooth, nearly elliptical cloud shapes, while the cyclones have large, filamentary, diffuse cloud patterns (**Fig. 4**). The large weather systems

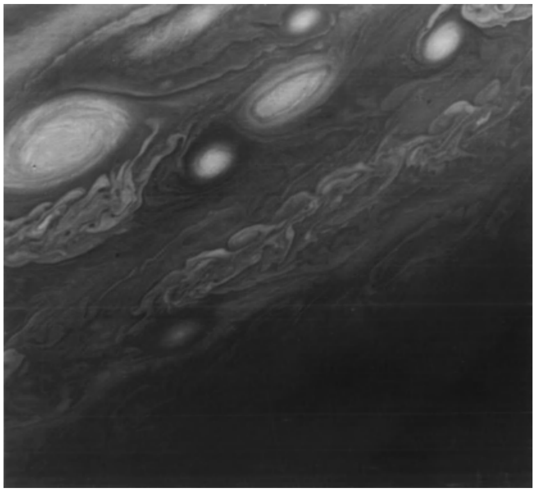


Fig. 4. Fine details of Jovian atmospheric features as photographed by *Voyager 1*. The dark halo surrounding the smaller bright spot to the right of large oval is a hot spot, a warmer region of Jupiter's atmosphere. (*NASA*)

move more slowly than the jets that constrain them.

Individual vortices form within the zones and belts, some persisting for decades. The major white ovals are hot-spot regions of strong infrared emission and are exceptionally dry regions of descending air, while other features such as the plumes observed in the equatorial region may be surface phenomena. The great scale changes in the belts and zones can be gradual (time scales of several months) or cataclysmic and sporadic (several days). For example, in 1990 a new temperate belt was created over a 3-month period as several bright features and white and dark spots coalesced.

Oscillations. Jupiter has global acoustic-mode radial oscillations, discovered in 1991, with periods of 8–17 min. Therefore, the entire planetary atmosphere also has cells of regular vertical motion which may contribute to the complexities of its meteorology.

Comet impact. During July 16–22, 1994, twenty-one observable fragments of Comet Shoemaker-Levy 9 impacted the Jovian atmosphere at nearly 120,000 mi/h (200,000 km/H), the first time such an event had been witnessed. Though occurring on Jupiter's nonvisible side, each strike was clearly visible as a dark area in the south temperate zone as the planet's rapid rotation brought it into view. In addition, ejecta and impact plumes were seen to rise approximately 1860 mi (3000 km) above the planetary limb. The impact of fragment G, thought to be the largest at 2–3 mi (3–4 km) in diameter, released energy equivalent to 6×10^{12} tons of TNT and left a dark spot 1550 mi (2500 km) in diameter surrounded by concentric rings (**Fig. 5**), which expanded at 0.28 mi/s (0.45 km/s). Each strike produced superheated gaseous bubbles, observed at infrared wavelengths, which rose above the stratospheric methane and cooled into aerosols (silicate dust, hyrocarbons, or hydrogen cyanide), as well as increased aurorae near the sites.

The event provided a first probe into the Jovian atmosphere. Traces of ammonia, sulfur (S_2), and hydrogen sulfide in the fireballs showed that they penetrated the upper two cloud layers of ammonia and ammonium hydrosulfide, but the absence of water vapor suggests that the comet itself was low in oxygen-rich ices and volatiles. Surges in Jupiter's decimetric radio emission during the impacts suggested that large numbers of electrons were injected into the Jovian magnetosphere. Given the number of such objects in the solar system, it is thought that Jupiter is impacted once every 500 years by a 0.2-mi-diameter (0.3-km) object and every 6000 years by a 1-mi (1.6-km) comet such as Shoemaker-Levy 9. *See* COMET.

Interior composition and structure. Jupiter primarily consists of liquid and metallic hydrogen. Early measures of the ratios of helium, carbon, and nitrogen to hydrogen gave values resembling those of the Sun, and therefore the primordial composition of the solar system. However, later analyses of the methane spectrum showed a two- to threefold overabundance of carbon as compared to solar values, a result confirmed by gravity analyses of the rocky core. Still in a late phase of its gravitational

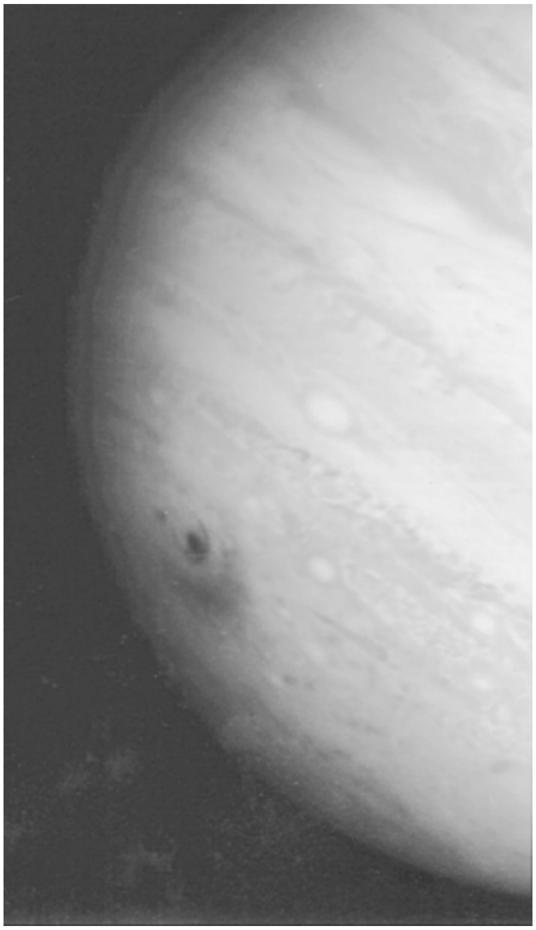


Fig. 5. Image of Jupiter from the *Hubble Space Telescope* showing impact sites of fragments D and G of Comet Shoemaker-Levy 9. The large feature with a ring and crescent-shaped ejecta was created by fragment G. The smaller feature to the left of fragment G impact site was created by fragment D. (*H. Hammel and NASA*)

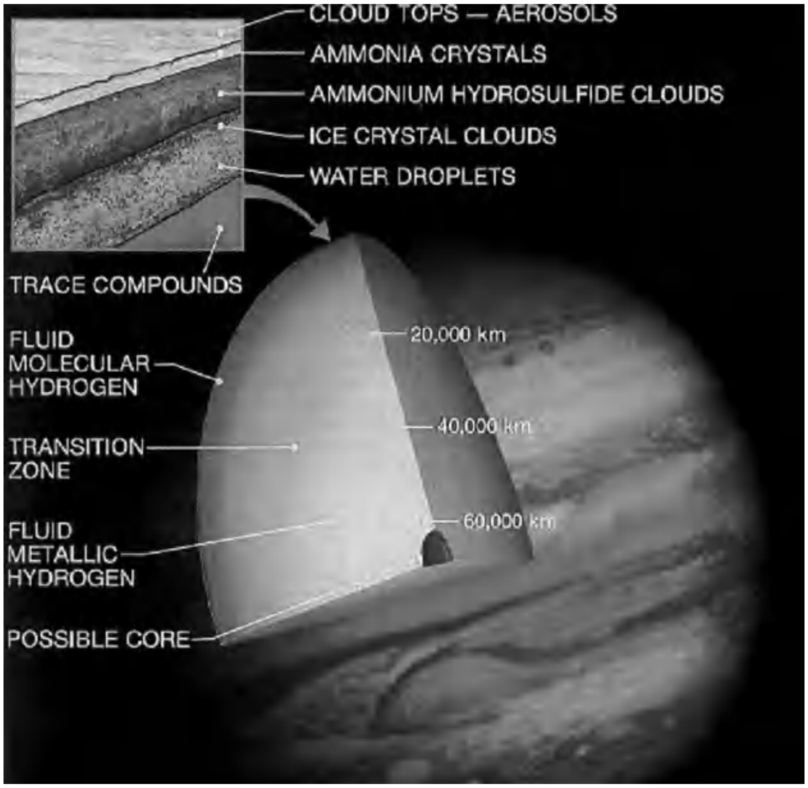


Fig. 6. Schematic diagram of the interior of Jupiter.

contraction, the planet converts the released gravitational energy into heat, emitting 1.668 times as much thermal energy as it receives from the Sun.

Below the clouds is a 13,000-mi-thick (21,000-km) layer of hydrogen and helium, which changes from gas to liquid as the pressure increases. Beneath the liquid hydrogen layer is a 25,000-mi-deep (40,000-km) sea of metallic hydrogen, where the pressure is approximately 3 million times that at Earth's sea level. This layer is electrically conductive and causes Jupiter's intense magnetic field. Acoustic oscillations confirm the presence of a small high-density solid core, perhaps 1.5 times the Earth's diameter, with a density 10–30 times that of Earth's mean density and a temperature of 55,000°F (30,000°C) (**Fig. 6**).

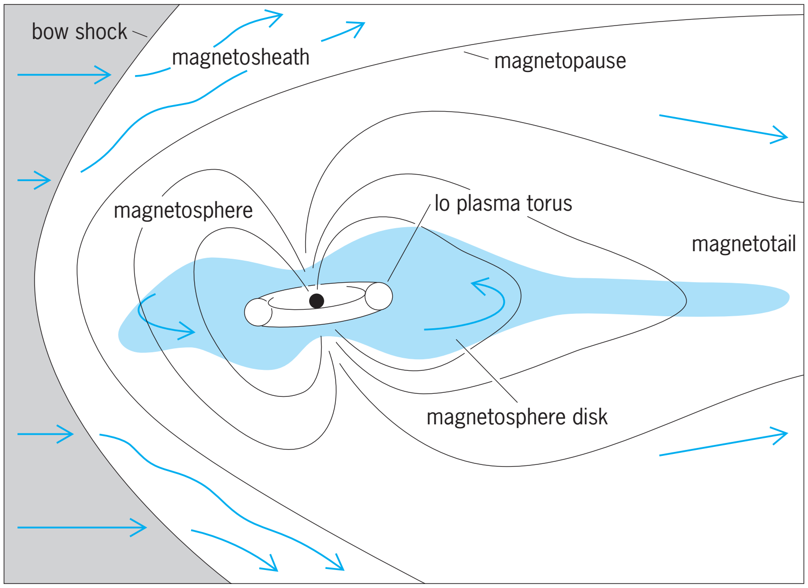


Fig. 7. Jovian magnetic environment. (*From Return to Jupiter, Astronomy, 7(9):6–24, 1979*)

Differentiation occurs. The metallic envelope has a mass fraction of heavy elements that is less than 7.5 times the solar value, while this value ranges from 1 to 7.2 in the molecular envelope. The *Galileo* probe found that the helium-to-hydrogen ratio was 0.156 (near solar), but that carbon, sulfur, and xenon were enhanced and oxygen and neon depressed with respect to solar values. The large abundance of carbon suggests that gases other than those from the solar nebula contributed to the composition of Jovian volatile gases. Probably the entire planet holds 11–45 earth masses of elements other than hydrogen and helium.

Magnetosphere. Jupiter possesses the strongest magnetic field and most complex magnetosphere of any planet in the solar system (**Fig. 7**). The Jovian magnetosphere is the largest object in the solar system, extending well beyond the orbit of Saturn, which sometimes passes through it. If it could be seen visually from Earth, it would subtend several times the diameter of the full moon on the sky. The Jovian magnetic field rotates with the rotational period of the planet and contains an embedded plasma trapped in the field. At the distance of the satellite Io, the field revolves faster than the satellite, and so numerous collisions occur with the atmospheric gas of that body, resulting in the stripping away of 10^{28}–10^{29} ions per second. The energy involved slows the magnetic field, and so, beyond Io, the magnetic field no longer rotates synchronously with the planet. The ions removed from Io spiral around the magnetic lines of force, oscillating above and below the plane of Io's orbit. This ring of ions is known as the Io plasma torus and emits strongly in the ultraviolet. The motion of Io through the magnetosphere creates a 400,000-V, 2×10^{12} W circuit, sufficient to cause pronounced aurorae in both the equatorial and polar regions of the satellite.

Except near the planet, the major component of the Jovian magnetic field is dipolar and opposite in direction from that of the Earth. At a distance of three Jovian radii, the field strength is 0.16 gauss (16 microtesla). Closer to the planet, the field is quadrupolar and octopolar. There the field strength varies from 3 to 14 gauss (0.3 to 1.4 millitesla). The magnetic axis is aligned about 10° with respect to the rotational axis. The inner radiation ring flux intensity peaks at 2.2 Jovian radii, while its innermost edge is at 1.35 Jovian radii, well above the atmospheric cut-off. Radiation is also absorbed near the outer edge of the bright dust ring.

Ion sources for the magnetospheric plasma are the solar wind; the Jovian satellites, particularly Io; and the ionosphere itself. Ion temperatures in the Io plasma torus reach 17,500°F (10,000 K). The dayside Jovian magnetosphere extends to 105 Jovian radii. On the duskside high-latitude region, intense fluxes of counterstreaming ions and electrons from the edge of the plasma torus to the duskside magnetosphere are aligned tightly with the magnetic field, superimposed on a hot plasma background. In addition, a thin plasma sheet (its depth at 15 Jovian radii is 2 radii) has temperatures of 100–400 eV. At polar latitudes, energetically charged particles following the Jovian magnetic lines of force may have access to the interplanetary medium. The Jovian magnetotail extends at least 4.5 astronomical units (4×10^8 mi or 6.5×10^8 km). *See* SOLAR WIND.

Jupiter possesses the most powerful aurorae in the solar system. They are observed at soft x-ray, ultraviolet, visible, and infrared wavelengths. Energies are 1000 times higher than in terrestrial aurorae. Protons, electrons, sulfur, and oxygen ions (from Io) precipitate onto the planet near its magnetic poles. These also form H_3^+ (trihydrogen) ions, the first ever observed. The *Galileo* probes observed visible auroral rings, highly structured and rapidly varying in time, 300–5000 mi (500–8000 km) in diameter at an altitude of 150 mi (240 km) above the 1-bar pressure level. *See* AURORA; MAGNETOSPHERE.

Radio astronomy. Jupiter produces distinct types of radio emission. Thermal radiation from the high stratosphere is detectable at wavelengths below about 10 cm and indicates temperatures in the upper emitting layers of -280 to $-225°$F (100 to 130 K). Microwave nonthermal emission in the band from about 3 to 70 cm arises from synchrotron radiation from relativistic electrons in extended Jovian Van Allen belts. Decametric radio emission (of frequencies 10–23 MHz) is also formed by synchrotron emission of these electrons and is 100% elliptically polarized. It occasionally produces S-bursts, with durations of tens of milliseconds, caused by instabilities of the electron beam in the Io-Jupiter flux tube. Hectometric emission (around 500 kHz) is probably rooted in the auroral ovals. The *Ulysses* probe detected radio emission at 3–30 kHz. Some radio noise is thought to originate in the lightning storms first observed by *Voyager 2* at depths of 50 mi (80 km) below the ammonia clouds. *See* POLARIZATION OF WAVES; SYNCHROTRON RADIATION; VAN ALLEN RADIATION.

Subsequent observations by the *Galileo* space probe have shown that lightning occurs in the Jovian atmosphere at a rate of approximately 20 bolts per second, approximately one-tenth the Earth's rate. The storms producing this lightning have been observed to be over 600 mi (1000 km) in diameter and over 45 mi (75 km) tall. Most lightning storms are concentrated between north latitudes 47 and 49°.

Periodicities in the radio-noise storms and rocking of the polarization plane of the microwave nonthermal emission led to the well-determined radio rotation period 9h55m29.7s. The difference between the radio and the various other observed Jovian rotation periods suggests that the core of Jupiter is rotating about 13 s faster than the mantle and that angular momentum may be significantly exchanged among the core, mantle, and atmosphere over periods of years. *See* RADIO ASTRONOMY.

Jovian ring. *Voyager 1* and *2* first detected a faint ring encircling Jupiter. Later probes found that Jupiter possesses a tripartite ring system. There is a flat main ring, a lenticular toroidal halo interior to this, and a gossamer ring which is exterior to the main ring. The entire ring system begins at about 55,000 mi (92,000 km) from Jupiter's center and extends to 150,000 mi (250,000 km). The main ring has a radial structure with dips in its apparent brightness due to the perturbations of the closer satellites. It is largely made up of 0.3–1.6-micrometer particles with human-scale lifetimes, whose source is dust and debris knocked off the Jovian moons XV Adrastea and XVI Metis by impacts with meteoroids, comets, and asteroids. Its inner and outer radii are respectively 76,900 mi (123,000 km) and 80,590 mi (128,940 km). The estimated mass of all particles forming it is 1×10^{13} kg. The toroidal halo interior to this ring was probably formed by electromagnetic forces pushing the smallest grains out of the ring plane. The exterior gossamer ring has two components, one associated with V Amalthea and the other with XIV Thebe. The Amalthea ring extends to precisely the orbital distance of the satellite and ends abruptly, while extending to one-half the thickness of the moon's excursion off the Jovian equatorial plane. The external gossamer ring particles orbit in retrograde orbits. The Amalthea ring forms the largest and sparsest ring known in the solar system. The material appears to be a dark reddish soot, consistent with the release of small particles from the two moons.

Satellites. As of October 2003, sixty-four known satellites had been discovered to orbit Jupiter, the largest number for any planet in the solar system. The four largest are I Io, II Europa, III Ganymede, and IV Callisto, discovered by Galileo in 1610 in Italy and independently by Simon Marius in Germany, who named them (though they are collectively known as the Galilean satellites) [see **table**].

The four Galilean satellites are of fifth and sixth stellar magnitudes and would be visible to the naked eye if they were not so close to the much brighter parent planet. They are easily visible in binoculars. The majority of the others are very faint objects only recently discovered with large telescopes.

The planes of the orbits of the major satellites are inclined less than 0.5° to the equatorial plane of Jupiter, so that with the occasional exception of IV, they are eclipsed in Jupiter's shadows and transit in front of its disk near conjunction. The eclipses, transits, and occultations of Jupiter's satellites led to the discovery of the finite velocity of propagation of light by O. Roemer in 1675. Satellites VIII, IX, XI, and XII have retrograde motion. *See* ECLIPSE; OCCULTATION; TRANSIT (ASTRONOMY).

The four Galilean satellites are orbitally connected. It is thought that Io was once much closer to Jupiter. Over the course of 100 million years, rotational energy was transferred to the satellite, causing it to move outward until it locked Europa into a 2:1 synchronous orbit. The two satellites then spiraled out as a pair and changed the eccentricity of Ganymede's orbit. Eventually, the mutual synchronicity will begin

The 16 brightest satellites of Jupiter

Satellite	Diameter, mi (km)	Mean distance from center of planet, 10^3 mi (10^3 km)	Orbital period, days	Magnitude at mean opposition	Year of discovery	Rotation*	Mass, 10^{20} lb (10^{20} kg)
XVI Metis	37 × 21 (60 × 34)	80 (128)	0.294	17.5	1979	Sync.	0.002 (0.001)
XV Adrastea	12 × 9 (20 × 14)	80 (129)	0.297	18.7	1979	Sync.	0.0004 (0.0002)
V Amalthea	155 × 80 (250 × 128)	113 (181)	0.498	14.1	1892	Sync.	0.16 (0.07)
XIV Thebe	72 × 52 (116 × 84)	138 (222)	0.674	16.0	1979	Sync.	0.017 (0.008)
I Io	2256 (3630)	262 (422)	1.769	5.0	1610	Sync.	1971 (894)
II Europa	1950 (3138)	417 (671)	3.551	5.3	1610	Sync.	1058 (480)
III Ganymede	3270 (5262)	665 (1,070)	7.155	4.6	1610	Sync.	3267 (1482)
IV Callisto	2983 (4800)	1170 (1883)	16.689	5.6	1610	Sync.	2372 (1076)
XIII Leda	10 (16)	6893 (11,094)	240	19.5	1974	Unknown	0.00013 (0.00006)
VI Himalia	116 (186)	7133 (11,480)	251	14.6	1904	Nonsync.	0.21 (0.10)
X Lysithea	22 (36)	7282 (11,720)	260	18.3	1938	Nonsync.	0.0017 (0.0008)
VII Elara	47 (76)	7293 (11,737)	260	16.3	1905	Nonsync.	0.017 (0.008)
XII Ananke	19 (30)	13,173 (21,200)	631	18.8	1951	Nonsync.	0.0008 (0.0004)
XI Carme	25 (40)	14,043 22,600	692	17.6	1938	Unknown	0.002 (0.001)
VIII Pasiphae	31 (50)	14,602 23,500	735	17.0	1908	Unknown	0.004 (0.002)
IX Sinope	22 (36)	14,726 (23,700)	758	18.1	1914	Nonsync.	0.0017 (0.0008)

* Synchronous or nonsynchronous.
SOURCE: After Lunar and Planetary Laboratory, University of Arizona; and J. A. Wood, Forging the planets: The orgin of our solar system, *Sky Telesc.*, 97(1):36–48, January 1999.

to affect the orbit of Callisto and lock it in, too. The gravitational effects of proximity of Jupiter and the slight elliptical orbits cause surface flexure and tidal heating strong enough to produce internal melting and major upheaval. All four satellites revolve within the Jovian magnetosphere, causing heavy ion bombardment of their surfaces.

The Jovian satellites fall into eight groups, based on their orbital characteristics. The Small Inner Irregulars and Rings (XVI Metis, XV Adrastea, V Amaltheal, XIV Thebe) are closest to the planet and govern ring structure. Irregular refers to their origin: they are probable captures from the asteroid belt after Jupiter's formation. The Galilean moons (as above) undoubtedly formed independently from the solar nebula or else as agglomerations of nebular material, likely at the same time as Jupiter. The Themisto (XVIII) and Himalia (VI) Prograde Irregular Groups comprise five satellites in total, all with moderately high eccentricites. They, too are likely captures and orbit the planet in prograde orbits. Three other groups (Retrograde Irregular, Carme [XI] Retrograde Irregular, and Pasiphae [VIII] Retrograde Irregular) orbit Jupiter in a retrograde direction and have moderate eccentricities. The newest discoveries have not been assigned to a group, though they share similar orbits and may be fragments of a single large body that broke up. In appearance, they are very dark and quite faint.

Io. The close approaches of the *Voyager* and *Galileo* spacecraft as well as the superior imaging capabilities of the *Hubble Space Telescope* have shown the four Galilean satellites to be very different (**Fig. 6**). I Io is probably the most geologically active body in the solar system. Its surface landforms include active shield volcanoes, calderas, mountains, plateaus, flows, grabens, and scarps. There are strong concentrations of craters near the sub- and anti-Jovian points. High mountains are equally distributed over the surface and are not connected with volcanoes. The tallest mountain known has an altitude of 52,000 ft (16 km). High-temperature (1700–2000 K) magnesium-rich silicate volcanism is common. Volcanism is likely due to a slight eccentricity in the orbit produced by the proximity of II Europa, which causes internal tidal forces to produce heat due to friction. Over 80 active volcanoes have been observed, though there are over 200 calderas larger than 12 mi (20 km) in diameter; by contrast, Earth has but 15. The most active volcanoes are Loki (responsible for half the satellite's total heat output) and Pele.

In 1999 a stupendous eruption was seen at Tvashtar Patera volcano. The lava fountain was observed to be over 0.9 mi (1.5 km) tall and 25 mi (40 km) long. The lava produced was hotter than 2420°F (1600 K) and was visible to Earth-based telescopes observing in the infrared.

Volcanic eruptions generally produce plumes resembling geysers containing fine dust and cold SO_2 (sulfur dioxide) gas and snow with exit velocities of 500–1000 m/s. Other compounds such as diatomic sulfur (S_2) are also ejected. Once on the surface, S_2 changes to S_3 and S_4, which are bright red. Over time, these become S_8 (the common yellow form of sulfur). Therefore, the most active volcanoes are surrounded by crimson rings which fade with time. Other bright colors on the surface are white and gray (SO_2 frost), yellow and brown (sulfurous compounds), red and black (showing the most recent activity), and green (origin unknown). Since Io possesses an ionosphere, there are auroral glows near the volcanic plumes. It is thought that a global surface layer is deposited at a rate of 1 mm to 1 cm every year, thereby explaining the total lack of impact craters. The internal structure of Io is thought to consist of an iron and iron sulfide core, surrounded by a molten silicate interior and a 19-mi-thick-(30-km) silicate crust, able to support the mountains.

Io has been found to have an intrinsic magnetic field of an amplitude consistent with an internally driven dynamo. Therefore, it is a magnetized, solid planet. However, the field may turn on and off over million-year time scales as the interior recycles.

Despite its small size, Io also possesses a variable atmosphere consisting primarily of sulfur dioxide (with minor components of oxygen, sodium, and sulfur), which has been detected at infrared, millimeter, and ultraviolet wavelengths. It is extremely localized, covering only 10% of the satellite at any given time. Atmospheric pressure is only about 1 microbar. Recently it has been found that the proportion of chlorine in the Io atmosphere is the highest of any object in the solar system and may result from a volcanic eruption component of salt (NaCl), proceeding from underground rivers of salt water. The Jovian magnetosphere sweeps away 1000 kg of volcanic gases and surface material per second. This is the source of material in the Io plasma torus. The sodium cloud thus produced is as large as the Jovian magnetosphere itself, and has been detected not only in the Jovian vicinity but also in interplanetary space.

Europa. Europa possesses the highest albedo of any object in the solar system. The satellite has a varied surface terrain, dominated by tectonic deformation, impact cratering, and the possible emplacement of ice-rich materials and perhaps liquids on the surface. The leading hemisphere is covered with water frost, while the following is darker and redder, possibly caused by the bombardment of sulfur ions from Io. The underlying bright blue plains are the basic surface feature from which the other terrain types are derived. Tectonically deformed bright plains form the mottled regions through local disruptions that allow subsurface salts and clays to well up, akin to sea-floor spreading on Earth. There are striking linear features which range from several miles to 13 mi (20 km) across and which extend for 625 mi (1000 km) or more. These may demonstrate the existence of geyserlike cryovolcanism. Initial eruptions could involve gases or liquids plus rocks. Complex ridges also exist with heights of 330–660 ft (100–200 m), showing that parts of the icy crust have been modified by the intense faulting driven by energy from the planetary interior (**Fig. 9**). These ridges have a reddish coloration at their tops which may be from mineral contaminants in the ice. Finally, there are also impact craters with visible ejecta, though the centers are curiously muted, as though filled immediately with a slushy material. The age of the visible surface has been estimated to be less than 100 million years, due to the tectonic activity. The changing pattern of ridges is consistent with a nonsynchronous rotation of an ice shell with respect to a synchronously rotating interior. Dark patches may show venting of salt-laden water from below the ice. The core of Europa is thought to be iron-rich and is 781 mi (1250 km) in diameter. It is surrounded by a silicate mantle overlain by a subsurface ocean and a thin ice crust. The total ice shell thickness may be 6–94 mi (10–150 km), but it may thin to less than 0.6 mi (1 km) in some places. This is shown by visible "rafting": floating icebergs. That the salt-water ocean of Europa is still liquid below the surface is suggested by the observation that the slight magnetic field observed by *Galileo* changes direction to follow the Jovian magnetosphere. This could occur only in an induced field carried by a salt-rich true liquid layer. Europa also possesses a tenuous oxygen (O_2) atmosphere. It is not inconceivable that simple life could exist in the Europan oceans, much the same way as life-forms congregate near thermal vents in Earth's

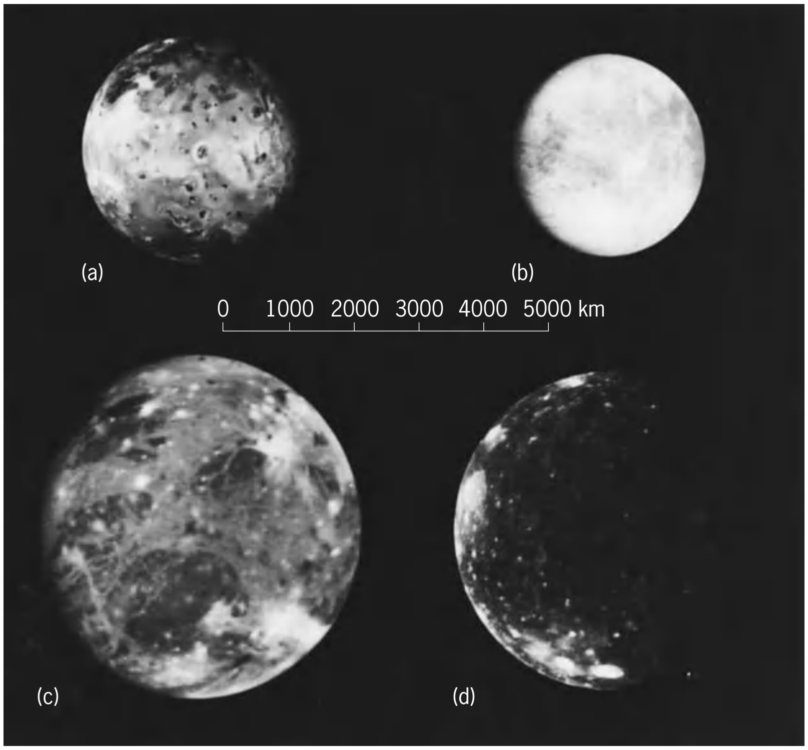


Fig. 8. ***Voyager 1*** **photographs of the Galilean satellites, shown to scale. (*a*) Io. (*b*) Europa. (*c*) Ganymede. (*d*) Callisto. 1 km = 0.6 mi. (*NASA*)**

Fig. 9. Intricately textured ridged plains on Europa, photographed by *Galileo* on December 16, 1997, from a distance of 800 mi (1300 km). The region covered is about 12 mi (20 km) on a side, with a resolution of only 85 ft (26 m) per picture element. Multiple generations of overlapping ridges are visible, many of them double. (*Jet Propulsion Laboratory*)

oceans. This possibility led *Galileo* project managers to choose to destroy the probe in 2003 by allowing it to fall into the Jovian atmosphere, rather than to contaminate a possible European ecosystem with Earth microbes.

Ganymede. The density of impact craters on the surface of III Ganymede shows that it is several billion years old. There are two major types of terrain: older, heavily fractured dark and rocky areas and brighter, less densely cratered areas rich in water and dry ice. Crossing the dark terrain are 3–6 mi-wide (5–10 km) furrows which may have formed from impacts which released water ice from below in a form of cryovolcanism. Dark maerial on the surface is likely meteoritic dust. Water and dry ice frost many of the impact craters. Fifty percent of the surface is made up of grooved terrain caused by tectonic stretching of the moon's surface into grabens. Some of the low-lying valleys were flooded with slush about 1 billion years ago, leading to some thought that a subsurface ocean may exist here, also. The partially molten iron core of Ganymede may take up to 50% of its diameter. This is surrounded by a silicate lower mantle, an icy upper mantle, and a thin ice crust. The satellite possesses a magnetic field with a strength 1% that of Earth's. It may possibly form in a salty, electrically conducting layer of water some 110 mi (170 km) below the surface.

Ganymede possesses a slight atomic hydrogen and O_2-O_3 (ozone) atmosphere both above the surface and trapped within the surface ice. In fact, ozone is surprisingly abundant near the poles, while the O_2 is found near the equator.

There is faint fluorescence of oxygen in the ultraviolet at the poles, forming an aurora, while there is a visible light aurora at the equator. A dusting of frost near the poles may derive from the manner in which Ganymede's magnetic field interacts with Jupiter's. Ions and electrons preferentially cascade over the polar regions, which may cause sublimation of water which then accumulates as piles of frost.

Callisto. Since it is similar in radius and density, IV Callisto should have a geological history resembling that of Ganymede, but does not. It is the most globally heavily cratered of the Galilean satellites, with large impact craters (for example Valhalla, with a diameter of 375 mi [600 km]) formed early in its history. The heavy cratering is indicative of its great age, but there is a curious lack of small craters, suggesting that there is ice-rich debris that subsumes them. Sublimation erosion by the Sun is the dominant process, causing crater rims and ridges to become heavily eroded. A long chain of craters is visible which was likely caused by the impact of a comet akin to Shoemaker-Levy 9. The centers of impact craters are often very bright, suggesting frost deposits. There is definite SO_2 frost concentrated on the forward-facing hemisphere. The impact craters can be muted, as if water ice slush filled them after meteoritic impact. There are few signs of internal geology. The crust may be 124 mi (200 km) thick, with a salty ocean 6 mi (10 km) thick beneath. Below this is compressed rock and ice, most forming an undifferentiated core. The liquid ocean is suggested by magnetic irregularities observed by Galileo; there is no intrinsic magnetic field, nor is there an atmosphere. This curious internal structure shows that Callisto formed in a cold environment, unlike the other three Galilean satellites which achieved differentiated cores by forming in a hot region.

Other satellites. The Galileo spacecraft was the first to image the four inner satellites of Jupiter, XVI Metis, XV Adrastea, V Amalthea, and XIV Thebe. All are 25–35% brighter on their leading sides than on their trailing sides. Metis and Adrastea are redder in coloration. Bright spots were seen on Amalthea and Thebe (diameters less than 13 mi [20 km]), as well as what appeared to be higher albedo patches on ridges and crater rims. These probably represent deposition of material from Io. The major geologic features on all four inner satellites are impact craters. The largest is 41 mi (60 km) across on Amalthea. Since its density is too low to allow it to be a solid object, Amalthea is almost certainly a loose agglomeration, held together by gravity. If it ever was solid, it may have been shattered into sand and boulder-sized rubble by impacts.

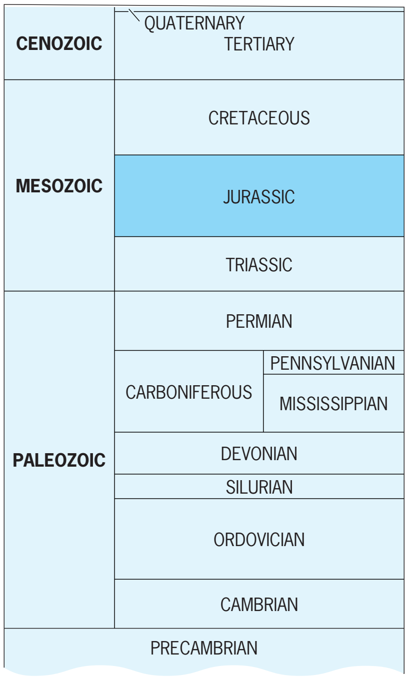
VI Himalia has been shown from photometry to consist of carbonaceous chondritic material, and probably the remainder of the satellites of Jupiter are constructed of similar material. Satellites VII–XII are asteroidal in nature. They resemble the Trojan asteroids in their colors and albedos, and may share a common origin. They are less elliptical in shape than the other small Jovian satellites, perhaps because they spent more time in the proto-Jupiter nebula. However, their properties are also consistent with a capture origin, and their orbits are subject to large perturbations by the Sun. Thus they may form part of a fluctuating population gained and lost over very long time spans. *See* ASTEROID; PLANETARY PHYSICS; TROJAN ASTEROIDS.

Elaine M. Halbedel

Bibliography. J. K. Beatty, *Galileo*: An image gallery II, *Sky Telesc.*, 93(3):28–35, March 1997; J. K. Beatty and D. H. Levy, Crashes to ashes: A comet's demise, *Sky Telesc.*, 90(4):18–26, October 1995; J. K. Beatty, C. C. Petersen, and A. Chaikin (eds.), *The New Solar System*, 4th ed., 1998; M. J. S. Belton (ed.), *In Orbit at Jupiter: Contributions of the Galileo Imaging Science Team*, NASA Galileo Proj. Doc. 625–801, 2000; M. J. S. Belton et al., Galileo's first images of Jupiter and the galilean satellites, *Science*, 274:377–385, 1996; D. Fischer, *Mission Jupiter: The Spectacular Journey of the Galileo Spacecraft*, Copernicus Books, 2001; M. Hanlon, *The Worlds of Galileo: The Inside Story of NASA's Mission to Jupiter*, St. Martin's Press, 2001; D. M. Harland, *Jupiter Odyssey: The Story of NASA's Galileo Mission*, Springer-Praxis, 2000.

Jurassic

The system of rocks deposited during the middle part of the Mesozoic Era, and encompassing an interval of time between about 200 and 142 million



years ago, based on radiometric dating. It takes its name from the Jura Mountains of Switzerland. Its rich marine invertebrate faunas in western Europe have been the subject of intensive study since the pioneering days of geology in the early nineteenth century, and provided the basis for the fundamental stratigraphic concepts of stages and biozones. *See* DATING METHODS.

Subdivisions. The Jurassic System is subdivided into 11 stages which, with the exception of the Tithonian, are named from localities in England, France, and Germany (**Fig. 1**). These, and the much greater number of zones, are based upon ammonites, which are by far the most valuable fossils biostratigraphically because of their high rate of species turnover in time due to rapid evolution and extinction. The most refined stratigraphic subdivisions have been made in the British Isles, with 54 zones and 176 subzones. Because of biogeographic provinciality, with different ammonite taxa inhabiting Boreal and Tethyan realms, difficulties of correlation can occur for younger Jurassic strata; and the youngest stage in the Tethyan Realm, the Tithonian, embracing most of the world, is equivalent to the Volgian stage of the Boreal Realm, extending from northern Eurasia to northern North America. The ammonites of the Volgian are quite different from those of the stratigraphically equivalent Tithonian. In the absence of ammonites, dinoflagellates are the most useful marine fossils for correlation, but in nonmarine strata problems of correlation are considerable, and stratigraphically less satisfactory pollen and spores have to be used. *See* STRATIGRAPHY.

SUBSYSTEM	STAGE	AGE, 10^6 years
		142
UPPER	Tithonian	
		149
	Kimmeridgian	
		155
	Oxfordian	
		157
MIDDLE	Callovian	
		160
	Bathonian	
		166
	Bajocian	
		174
	Aalenian	
		178
LOWER	Toarcian	
		183
	Pliensbachian	
		192
	Sinemurian	
		197
	Hettangian	
		200

Fig. 1. Succession of Jurassic stages, with estimated radiometric ages in millions of years. (*After J. Palfy et al., A U-Pb and ^{40}Ar/^{39}Ar time scale for the Jurassic, Can. J. Earth Sci., 37:923–944, 2000*)

Paleogeography and sea level. The main continental masses were grouped together as the supercontinent Pangaea, with a northern component, Laurasia, separated from a southern component, Gondwana, by a major seaway, Tethys, which expanded in width eastward (**Fig. 2**). From about Middle Jurassic times onward, this supercontinent began to split up, with a narrow ocean being created between eastern North America and northwestern Africa, corresponding to the central sector of the present Atlantic Ocean. At about the same time, and continuing into the Late Jurassic, separation began between the continents that now surround the Indian Ocean, namely Africa, India, Australia, and Antarctica. As North America moved westward, it collided with a number of oceanic islands in the eastern part of the Paleo-Pacific. Because the impingement was an oblique one, there was a general tendency for these accreted landmasses to be displaced northward along the cordilleran zone of the subcontinent. Other examples of so-called displaced terranes are known on the Asian side of the North Pacific, and some of the accretion of oceanic islands took place in Jurassic times.

There were also important paleogeographic changes later in the period involving the Tethys zone. An older, so-called Palaeotethys was progressively closed as an extensive, narrow continent known as Cimmeria, extending east-west, and collided with the southern margin of Eurasia. The name comes from the Crimean Peninsula of Russia, where there is well-displayed evidence of an intra-Jurassic orogenic disturbance indicative of continental collision. *See* OROGENY; PALEOGEOGRAPHY.

Sea level rose progressively through the period, with a corresponding flooding of the continents by shallow epeiric seas, that is, shallow seas that covered part of the continents but remained connected to the ocean. At the beginning, such seas covered less than 5% of the continents, but near the end, in Oxfordian and Kimmeridgian times, they covered approximately 25% (Fig. 2). The Jurassic sea-level curve also shows a succession of smaller-scale changes, of a duration of a few million years. Some of these, such

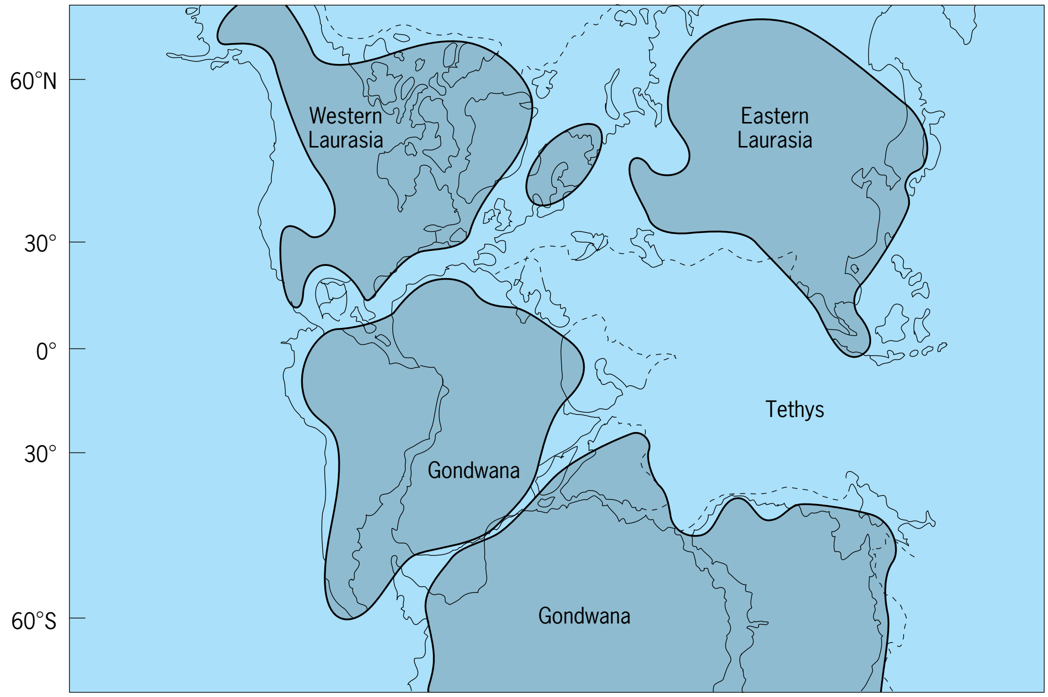


Fig. 2. Approximate distribution of land and sea in the Oxfordian stage. Small islands are excluded, but boundaries of modern continents are included as a reference.

as the Early Toarcian sea-level rise, are clearly global or eustatic, but others are more controversial and may reflect regional tectonic activity rather than truly global phenomena. It is uncertain by how much the sea level rose during the course of the period; but by using a hypsometric method, an estimate of between 330 and 500 ft (100 and 150 m) can be made. *See* PALEOCEANOGRAPHY.

Climate. The climate of Jurassic times was clearly more equable than at present, as indicated by two sets of facts. The first concerns the distribution of fossil organisms. Thus a number of ferns whose living relatives cannot tolerate frost are distributed over a wide range of paleolatitudes, sometimes as far as 60° N and S. Similarly, coral reefs, which are at present confined to the tropics, occur in Jurassic strata in western and central Europe, beyond the paleotropical zone. Many other groups of organisms had wide latitudinal distribution, and there was much less endemism (restriction to a particular area) with respect to latitude than there is today. The second set of facts concerns the lack of evidence for polar icecaps, such as extensive tillites or striated pavements.

However, there must have been strong seasonal contrasts of temperature within the Pangean supercontinent, and climatic modeling suggests winter temperatures at zero Celsius at or close to the paleopoles. A limited amount of evidence from northern Siberia and arctic North America, in the form of apparent glacial dropstones and glendonites, suggests the possibility of some ice, but this ice is likely to have been seasonally transient and small in volume.

There is no evidence of any significant change in the temperature regime through the Jurassic, but there are indications of a change in the humidity-aridity spectrum. Unlike the present, there were no tropical rainforests. Instead, a large area of western Pangea experienced an arid to semiarid climate in low latitudes, especially at some distance from the ocean. Precipitation is likely to have been dominantly monsoonal rather than zonal, a pattern unlike that of today. For most of the period, the continental area represented today by Eurasia had a comparatively humid climate, as indicated in nonmarine sediments by coals and the abundance of the clay mineral kaolinite. Toward the end of the Jurassic, however, there was a change to a more arid climate, indicated by the disappearance of coals and kaolinite and the occurrence of evaporites such as rock salt and gypsum. The reason for this change is unclear, but it may be bound up with a rainshadow effect created by the collision of the Cimmerian continent. *See* PALEOCLIMATOLOGY; SALINE EVAPORITES.

Tectonics and volcanicity. Most of Pangea experienced tensional tectonics as the supercontinent began to break up. This is manifested by graben and half-graben structures, with associated alkaline volcanicity. By far the largest flood basalt province is that of the Karoo in South Africa, most of the basalts and associated igneous rocks being erupted in the Early Jurassic, prior to the breakup of Africa, Madagascar, and India. The Middle Jurassic Ferrar dolerites of Victoria Land, Antarctica, and the contemporaneous Tasmanian dolerites are further manifestations of tensional tectonics, as are earliest Jurassic basalts in eastern North America and Morocco, again signifying tension prior to the Atlantic opening. The North Sea region of western Europe is in effect an aborted oceanic rift, with a major phase of tensional activity and associated volcanicity in the Middle and Late

Jurassic. This did not lead, however, to the creation of true ocean. *See* BASALT; GRABEN.

Compressional tectonics associated with subduction of ocean floor took place in many parts of the Pacific margins, with associated calc-alkaline volcanicity. An excellent example is the Andes. The North Pacific margins were also associated with significant strike-slip faulting bound up with the accretion of displaced terranes. The other important zone of compressional tectonics was along the southern margin of Eurasia, and is involved with the collision of the Cimmerian continent. *See* FAULT AND FAULT STRUCTURES.

Since it is not plausible to invoke the melting and freezing of polar ice caps to account for Jurassic sea-level change, this change must be bound up with tectonic activity. The most plausible mechanism for accounting for long-term sea-level rise is the growth of oceanic ridges, displacing seawater onto the continents, but the cause of short-term sea-level changes is more obscure and remains controversial. *See* CONTINENTS, EVOLUTION OF; GEOSYNCLINE; MID-OCEANIC RIDGE; PLATE TECTONICS; SUBDUCTION ZONES.

Vertebrate fauna. The vertebrate terrestrial life of the Jurassic Period was dominated by the reptiles. The dinosaurs had first appeared late in the Triassic from a thecodont stock, which also gave rise to pterosaurs and, later, birds. From small bipedal animals such as *Coelophysis*, there evolved huge, spectacular creatures. These include the herbivorous *Apatosaurus*, *Brontosaurus*, *Brachiosaurus*, *Diplodocus*, and *Stegosaurus* as well as the carnivorous, bipedal *Allosaurus*. Only two rich dinosaur faunas are known from Jurassic deposits, the Morrison Formation of the United States Western Interior and the approximately contemporary Tendaguru Beds of Tanzania. The two faunas are strikingly similar at family and generic level, which strongly suggests that free land communications existed between western North America and East Africa until quite late in the period, a fact that is not easy to reconcile with some paleogeographic reconstructions. *See* DINOSAUR.

Flying animals include the truly reptilian pterosaurs and the first animals that could be called birds as distinct from reptiles, as represented by the pigeon-sized *Archaeopteryx*. There were two important groups of reptiles that lived in the sea, the dolphinlike ichthyosaurs and the long-necked plesiosaurs. Both of these groups had streamlined bodies and limbs beautifully adapted to marine life. Turtles and crocodiles are also found as fossils in Jurassic deposits. *See* ARCHAEORNITHES; PTEROSAURIA.

Jurassic mammals, known mainly from their teeth alone, were small and obviously did not compete directly with the dinosaurs. They included a number of biologically primitive groups such as the triconodonts, docodonts and multituberculates. The fish faunas were dominated by the holosteans, characterized by heavy rhombic scales. Their evolutionary successors, the teleosts, probably appeared shortly before the end of the period. *See* DOCODONTA; HOLOSTEI; MULTITUBERCULATA; TELEOSTEI; TRICONODONTA.

Invertebrate fauna. Because they are far more abundant, the invertebrate fossil faunas of the sea are of more importance to stratigraphers and paleoecologists than are the vertebrates. By far the most useful for stratigraphic correlation are the ammonites, a group of fossil mollusks related to squids. They were swimmers that lived in the open sea, only rarely braving the fluctuating salinity and temperature of inshore waters. They are characteristically more abundant in marine shales and associated fine-grained limestones. From a solitary family that recovered from near extinction at the close of the Triassic, there radiated an enormous diversity of genera. Many of these were worldwide in distribution, but increasingly throughout the period these was a geographic differentiation into two major realms. The Boreal Realm occupied a northern region embracing the Arctic, northern Europe, and northern North America. The Tethyan Realm, with more diverse faunas, occupied the rest of the world. *See* LIMESTONE; SHALE.

In most facies the bivalves, which flourished in and on shallow, muddy sea bottoms, are the most abundant and diverse of the macrofauna. They included many cemented forms such as *Ostrea*, recliners such as *Gryphaea*, swimmers such as the pectinids and limids, and rock borers such as *Lithophaga*. However, the majority were burrowers: either relatively mobile, shallow burrowers or forms occupying deep permanent burrows and normally still found in their positions of growth. *See* BIVALVIA; FACIES (GEOLOGY).

Brachiopods were much more abundant and diverse than they are today. The range of depths below the sea surface that they occupied is far wider than for the bivalves, and a definite depth zonation can be established in Europe, just as with the ammonites. *See* BRACHIOPODA.

Echinoderms are best represented as fossils by the crinoids and echinoids, and were all inhabitants of shallow seas, unlike some of the modern representatives of this class. The echinoids include both primitive regular forms, such as the cidaroids, and irregular forms, such as *Clypeus* and *Pygaster*. *See* ECHINODERMATA; PYGASTEROIDA.

Corals belonged to the still extant Scleractinia group and included reef builders such as *Isastrea* and *Thamnasteria*. Calcareous and siliceous sponges are also common locally, even forming reefs. It seems likely that the siliceous sponges inhabited somewhat deeper water than the corals. *See* SCLERACTINIA; SCLEROSPONGIAE.

The invertebrate microfaunas are represented by abundant foraminifera, ostracods, and radiolaria. Foraminifera and ostracods are of great value to oil companies in correlation studies. *See* OSTRACODA; RADIOLARIA.

Not all Jurassic invertebrates lived in the sea. Some lived in continental environments such as lakes and rivers; they include a few genera of bivalves, gastropods, and arthropods. These faunas are far less diverse than their marine counterparts. *See* ARTHROPODA; GASTROPODA; PALEONTOLOGY.

Flora. With regard to the plant kingdom, the Jurassic might well be called the age of gymnosperms, the nonflowering "naked seed" plants, forests of which covered much of the land. They included the conifers, gingkos, and their relatives, the cycads. Ferns and horsetails made up much of the remainder of the land flora. These and others of the Jurassic flora are still extant in much the same forms. *See* CYCADALES; GINKGOALES.

Remains of calcareous algae are widely preserved in limestone. Besides the laminated sedimentary structures produced by what have traditionally been regarded as blue-green algae but are actually cyanobacteria, and known as oncolites and stromatolites, there are skeletal secretions of other groups. Some of these are benthic forms, but many pelagic limestones are seen under the electron microscope to be composed largely of tiny plates of calcite, known as coccoliths, which are secreted by certain planktonic algae also called coccoliths. *See* ALGAE; CYANOBACTERIA; STROMATOLITE.

It seems likely that the Late Jurassic saw the emergence of the flowering plants, the angiosperms, since well-developed forms of this group existed in the Early Cretaceous. However, it is not quite understood how they emerged, and a satisfactory direct evolutionary ancestor has yet to be identified with certainty.

Economic geology. Jurassic source rocks in the form of organic-rich marine shale and associated rocks contain a significant proportion of the world's petroleum reserves. A familiar example is the Upper Jurassic Kimmeridge Clay of the North Sea, and its stratigraphic equivalents in western Siberia. Some of the source rocks of the greatest petroleum field of all, in the Middle East, are also of Late Jurassic age. *See* MESOZOIC; PETROLEUM GEOLOGY. A. Hallam

Bibliography. W. J. Arkell, *Jurassic Geology of the World*, 1956; J. W. C. Cope et al. *Jurassic*, pts. 1 and 2, Geol. Soc. Lond. Spec. Rep. 14 and 15, 1980; A. Hallam, Jurassic climates as inferred from the sedimentary and fossil record, *Phil. Trans. Roy. Soc. Lond.*, B341:287–296, 1993; A. Hallam, *Jurassic Environments*, 1975; A. Hallam, A review of the broad pattern of Jurassic sea-level changes and their possible causes in the light of current knowledge, *Palaeogeog., Palaeoclimatol., Palaeoecol.*, 167:23–37, 2001.

Jute

A natural fiber obtained from two Asiatic species, *Corchorus capsularis* and *C. olitorius*, of the plant family Tiliaceae (**Fig. 1**). These are tall, slender, half-shrubby annuals, 8–12 ft (2.5–3.5 m) tall. *See* MALVALES.

When harvested the stems are retted in tanks or pools to rot out the softer tissues. The strands of jute fiber are then loosened by beating the stems on the surface of the water. The fibers are not very strong and deteriorate quickly in the presence of moisture, especially salt water. Despite these weaknesses, jute is much used. It is inexpensive and easily spun and converted into coarse fabrics. It is made into gunny, burlap bags, sacks for wool, potato sacks, covers for cotton bales, twine, carpets, rug cushions, curtains, and a linoleum base. It is also used in making coarse, cheap fabrics, such as novelty dress goods. Most of the commercial supply comes from plants grown in the Ganges and Brahmaputra valleys in Bangladesh and India. *See* FIBER CROPS; NATURAL FIBER. Elton G. Nelson

Fig. 1. Morphological features of *Corchorus capsularis*.

A number of diseases that affect jute cause losses in yield and reduce fiber quality. "Runner" and "specky" fiber are primarily due to disease-producing organisms.

The fungus *Macrophomina phaseolina* is believed to cause the most serious disease of the two species of jute. It is seed-borne and soil-borne, and pycnidiospores from susceptible plants besides jute also serve as sources of infection. The stem, leaves, and roots of both young and older plants are subject to attack. Stem infection usually takes place through a leaf petiole or at a node (**Fig. 2**). Root rot is complicated in that severity is increased when *M. phaseolina* is in combination with other fungi, bacteria, or nematodes, such as *Fusarium solani, Pseudomonas* sp., and *Meloidogyne incognita*, respectively. *See* AGONOMYCETES; LEAF; NEMATA (NEMATODA); ROOT (BOTANY); SEED; STEM.

In contrast to *Macrophomina phaseolina*,

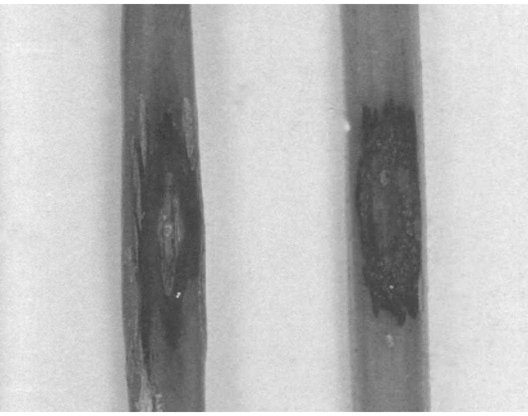


Fig. 2. Disease lesions on jute stems.

Colletotrichum capsici causes lesions on the stem internodes and may also attack seedlings and capsules of *C. capsularis. Macrophoma corchori* and *Diplodia corchori* cause stem diseases. Two species of bacteria, *Xanthomonas makatae* and *X. makatae* var. *olitorii*, attack the stem and leaves of both *C. capsularis* and *C. olitorius. Pythium splendens* causes a root rot and subsequent wilt of *C. capsularis*, and indications are that other species of *Pythium* also are root pathogens of jute. *See* FRUIT.

Other fungi which attack jute are *Sclerotium rolfsii*, *Curvularia subulata*, *Cercospora corchori*, *Rhizoctonia solani*, *Helminthosporium* sp., and *Alternaria* sp.

Seed treatments with approved fungicides are recommended for control of seed-borne pathogens and seedling diseases. Stem rot may be prevented by spraying with fixed copper compounds. The excessive use of nitrogenous fertilizers increases the incidence of stem diseases. Root-rot control requires the use of crop rotation and, in some areas, the use of recently developed varieties of *C. capsularis* which are more tolerant of certain root-rot pathogens than *C. olitorius*. *See* PLANT PATHOLOGY.

Thomas E. Summers

K

Kalahari Desert — Kyanite

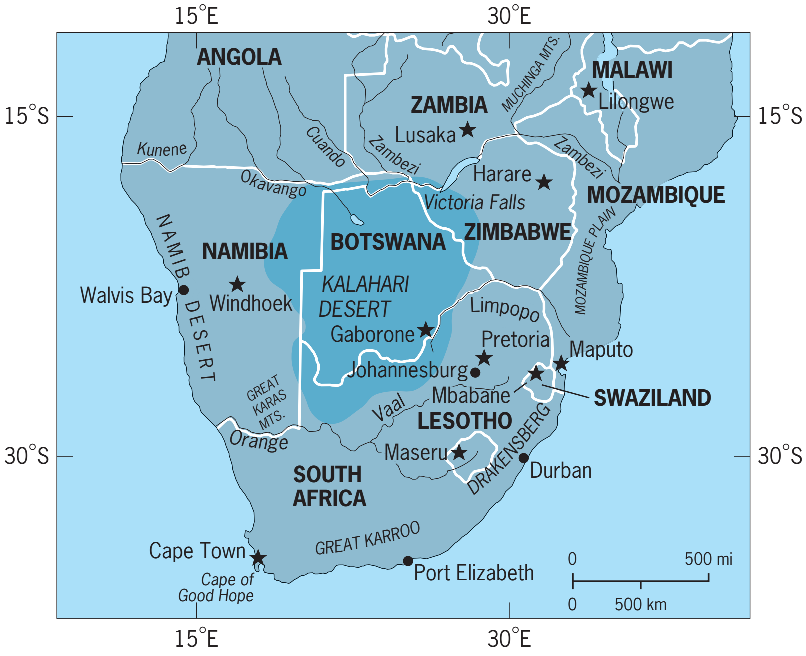
Kalahari Desert

A flat, largely waterless, sparsely populated, and sand-covered region in south-central Africa which occupies much of Botswana, plus adjacent parts of South Africa and Namibia. The precise boundaries of the Kalahari are difficult to define since it merges into humid regions to the north and east, and the arid Karoo and Namib further south and west. However, the generally agreed-upon boundaries are the Okavango-Zambezi "swamp zone" in the north and the Orange River in the south, with the western and eastern boundaries (in eastern Namibia and eastern Botswana, respectively) coinciding with the extent of the sandy Kalahari group sediments (see **illustration**). *See* AFRICA.

Climate in the Kalahari varies from arid to semiarid along a southwest–northeast gradient. The driest, southwestern areas (which can be considered true desert) receive as little as 150 mm (6 in.) of rainfall per year, with precipitation levels rising to the north and east to reach around 650 mm (26 in.) in northeast Botswana. However, much of this rain falls during the Southern Hemisphere summer months when high evapotranspiration rates also occur (exceeding 4000 mm or 157 in. per year in the southwest Kalahari), contributing to the desertlike conditions. *See* DESERT.

The Kalahari terrain is relatively flat, broken only by low hills along its periphery and dunefields in the extreme southwest. Permanent drainage is restricted to the north where the Chobe, Zambezi, and Okavango rivers dissect the landscape, the latter ending abruptly in the Okavango Delta swamps. Water is also present on a seasonal basis within the salt lakes of the Makgadikgadi and Etosha basins of central Botswana and northern Namibia, respectively, and within numerous smaller landscape depressions (called pans). In addition to rivers and lake basins, extensive networks of "fossil" dry valleys also cut across the region, suggesting that wetter climates may have occurred in the recent geologic past. *See* DUNE.

Soils developed within the sandy Kalahari sediments are typically skeletal (weakly bound) and have a relatively low nutrient status, although finer-textured soils may be found in association with rivers, pans, dry valleys, and interdune areas. Vegetation patterns broadly follow the climatic gradient, in terms of diversity and biomass, although, again, local variations occur near drainage features. Much of the region is covered by savanna communities, with shrub savanna prevalent in the southwest and tree savanna in the central and northern Kalahari. Exceptions occur in the Okavango swamps, where aquatic grasslands and riparian forests dominate, around saline lake complexes where savanna grasslands are common, and in the relatively wet north and east where dry deciduous forests are the



Map of the Kalahari Desert.

dominant vegetation type. *See* PLANT GEOGRAPHY; SAVANNA; SOIL.

The Kalahari has supported, and in some areas still supports, large populations of migratory wild herbivores, including antelope, zebra, and giraffe, along with associated predators, including lion, leopard, cheetah, and hyena. Pressures from hunting and increasing numbers of domestic herbivores, particularly cattle (with the associated fenced-in land), have meant that wildlife populations have been increasingly restricted to protected areas. At the same time, human populations are growing, most notably along the eastern and southern desert periphery (such as Gaborone, Botswana and Upington, South Africa), as well as within the desert core in conjunction with tourism (such as Maun, Botswana) and diamond mining (such as Orapa and Jwaneng, Botswana) developments. David J. Nash

Bibliography. A. S. Goudie, *Great Warm Deserts of the World*, Oxford University Press, 2002; D. Sporton and D. S. G. Thomas (eds.), *Sustainable Livelihoods in Kalahari Environments: Contributions to Global Debates*, Oxford University Press, 2002; D. S. G. Thomas and P. A. Shaw, *The Kalahari Environment*, Cambridge University Press, 1991; L. van der Post, *The Lost World of the Kalahari*, Penguin, London, 1962; E. N. Wilmsen, *Land Filled with Flies: A Political Economy of the Kalahari*, University of Chicago Press, 1989.

Kale

Either of two cool-season biennial crucifers, *Brassica oleracea* var. *acephala* and *B. fimbriata*, of Mediterranean origin. Kale belongs to the plant order Capparales, and the plant family Brassicaceae, commonly known as the mustard family. Cabbage, kohlrabi, broccoli, collards, and Brussels sprouts are all variants of *B. oleracea*

Kale is grown for its nutritious green curled leaves which are cooked as a vegetable (see **illus.**). Distinct varieties (cultivars) are produced in Europe for stock feed. Kale and collards differ only in the form of their leaves; both are minor vegetables in the United States. Cultural practices are similar to those used for cabbage, but kale is more sensitive to high temperatures. Strains of the Scotch and Siberian varieties are most popular. Kale is moderately tolerant of acid soils. Monthly mean temperatures below 70°F (21°C) favor best growth. Harvesting is generally 2–3 months after planting. Virginia is an important producing state. *See* BROCCOLI; BRUSSELS SPROUTS; CABBAGE; CAPPARALES; COLLARD; KOHLRABI. H. John Carew

Kale (*Brassica oleracea* var. *acephala*), cultivar Vates. (*Joseph Harris Co., Rochester, New York*)

Kaliophilite

A rare mineral tectosilicate found in volcanic rocks high in potassium and low in silica. Kaliophilite is one of three polymorphic forms of $KAlSiO_4$; the others are the rare mineral kalsilite and an orthorhombic phase formed artificially at about 930°F (500°C). It crystallizes in the hexagonal system in prismatic crystals with poor prismatic and basal cleavage. The hardness is 6 on Mohs scale, and the specific gravity is 2.61. At high temperatures a complete solid-solution series exists between $KAlSiO_4$ and $NaAlSiO_4$, but at low temperatures the series is incomplete. The principal occurrence of kaliophilite is at Monte Somma, Italy. *See* SILICATE MINERALS. Cornelius S. Hurlbut, Jr.

Kalsilite

A rare mineral described in 1942 from volcanic rocks at Mafuru, in southwest Uganda. Kalsilite has since been synthesized. It is one of the three polymorphic forms of $KAlSiO_4$; the others are kaliophilite and an orthorhombic phase formed artificially at about 930°F (500°C). The mineral as shown by x-ray photographs is hexagonal. The specific gravity is 2.59. In index of refraction and general appearance in thin section it resembles nepheline and is difficult to distinguish from it. Structurally the two minerals are similar but belong to different crystal classes. The rock in which kalsilite was found has a dark-colored, fine-grained matrix with large olivine crystals and greenish-yellow patches. These patches are intimate mixtures of diopside, calcite, and kalsilite. *See* FELDSPATHOID. Cornelius S. Hurlbut, Jr.

Kangaroo

Common name for a member of the family Macropodidae, the largest living marsupials. The 61 species in this family also include the wallabies, wallaroos, quokka, and pademelons. In Australia, the term kangaroo is applied only to the largest species of kangaroos, while all small and medium-sized ground-dwelling kangaroos are called wallabies. Except for their size, wallabies are practically the same as kangaroos in appearance as well as habits. *See* AUSTRALIA; MARSUPIALIA.

Description. There are many different kinds of kangaroos, ranging from the familiar great kangaroos

Red kangaroo (*Macropus rufus*). (*Photo by Gerald and Buff Corsi; © California Academy of Sciences*)

(*Macropus*) to the musky rat kangaroo (*Potoroidae*), which is roughly the size of a rabbit but more like a rat in appearance. There are three species of great kangaroos, all of which have extremely long, powerful, highly modified hindlimbs and relatively small, unspecialized forelimbs. The long, thick, muscular tail serves as a balancing organ and also adds impetus to the leaps. Large kangaroos can sustain hopping speeds faster than 55 km/h (35 mi/h). The western gray kangaroo or Forester (*M. fuliginosus*) has a brownish coat with black hands and feet. The eastern gray kangaroo (*M. giganteus*) is the only large kangaroo in Tasmania. It is more reddish-brown than the western gray kangaroo and has coarser and longer fur. The red kangaroo (*M. rufus*) [see **illustration**] is the largest and most gregarious kangaroo. It is richly colored, powerful, and graceful. The male is a brilliant wine-red. The female, or doe, is more slender than the male, is more lightly built, and has great speed. A smoky-blue in color, the female has been named the blue flyer. Adult males, or boomers, may have a head and body length of 1300–1600 mm (50–62 in.) and a tail length of 1000–1200 mm (39–46 in.). They may weigh 20–90 kg (43–197 lb) but seldom exceed 55 kg (120 lb). Females are smaller.

Kangaroos and their relatives primarily inhabit the prairies, open forest, and brush country in Australia but also are found in Tasmania and New Guinea. They feed mostly on plant foods including grasses, forbs (herbs other than grass), leaves, seeds, fruits, and bulbs. The dental formula is I 3/1, C O/O, Pm 2/2, M 4/4 for a total of 32 teeth. Following a gestation of 30–39 days, a single young (joey) is born. A newborn macropod is only 5–15 mm (0.2–0.6 in.) long with undeveloped eyes, hindlimbs, and tail. It uses its strong, well-developed forelimbs to crawl up and into its mother's pouch (marsupium), where it attaches to a teat. Depending on the species, the joey may remain completely in the pouch for 60–120 days. Following the pouch period, it will venture out to find food but will regularly return to the pouch, especially in times of danger. This period, which may last for several weeks to several months, is known as the young-at-foot period. When entering the pouch, the joey enters headfirst and then turns around and pokes its head back out. The joey is forced from the pouch just before the birth of the next young, but it can still put its head back into the pouch and suckle from the teat. The female is capable of producing different qualities of milk from the two teats—a feat achieved by having the mammary glands under separate hormonal control. Longevity in the wild is normally 12–18 years. Although great numbers of red and gray kangaroos are shot each year because they are considered pests of crops and pastures, the chief natural predator is the dingo (*Canis lupus familiaris*), the wild Australian dog.

Tree kangaroos. Tree kangaroos (*Dendrolagus*) are arboreal. They still retain the long hindlimbs for leaping, but the limbs have become shorter and broader, and the foot pads are rough to assist in climbing. They have a head and body length of 520–810 mm (20–32 in.) and a tail length of 408–935 mm (16–36 in.) They weigh 6.5–14.5 kg (14–32 lb). The pelage is usually long and may be blackish, brownish, grayish, or reddish. One species has a bright yellow face, belly, feet, and tail. The forelimbs and hindlimbs are nearly equal in size. The long, well-furred tail serves to help balance and steer the kangaroo as it makes long, flying leaps in the trees. The tail is also used to brace the animal when climbing. The 10 species live primarily in mountainous rainforest in Queensland (Australia) and New Guinea. They are very agile and are active both day and night. Food consists chiefly of tree leaves, ferns, and many kinds of fruits. Some species spend a considerable amount of time on the ground, whereas others spend almost all of their time in the trees. Clearing of rainforest and hunting are reducing their ranges. Habitat fragmentation may prevent dispersal and increase inbreeding. Three species are currently classified as endangered by the International Union for the Conservation of Nature and Natural Resources (IUCN), two are classified as vulnerable, and two as near-threatened.

Donald W. Linzey

Bibliography. B. Grzimek (ed.), *Grzimek's Encyclopedia of Mammals*, McGraw-Hill, 1990; D. Macdonald, *The Encyclopedia of Mammals*, Andromeda Oxford, 2001; R. N. Nowak, *Walker's Mammals of the World*, 6th ed., Johns Hopkins University Press, 1999.

Kaolinite

A common hydrous aluminum silicate mineral found in sediments, soils, hydrothermal deposits, and sedimentary rocks. It is a member of a group of clay minerals called the kaolin group minerals, which include dickite, halloysite (7 Å and 10 Å), nacrite, ordered kaolinite, and disordered kaolinite. These minerals have a theoretical chemical composition of 39.8% alumina, 46.3% silica, and 13.9% water [$Al_2Si_2O_5(OH)_4$], and they generally do not deviate from this ideal composition. They are sheet silicates comprising a single silica tetrahedral layer joined to a single alumina octahedral layer. Although the kaolin group minerals are chemically the same, each is structurally unique as a result of how these layers are stacked on top of one another. Kaolinite is the most common kaolin group mineral and is an important industrial commodity used in ceramics, paper coating and filler, paint, plastics, fiberglass, catalysts, and other specialty applications. *See* CLAY MINERALS; SILICATE MINERALS.

The term "kaolin" is derived from the Chinese word *Kau-ling* (high ridge), which is the name given to a hill in the Jiang Xi Province where white clay was mined for porcelain as early as the seventh century. Deposits of relatively pure kaolinite are called kaolin, which leads to some confusion because the terms are often used interchangeably; however, kaolinite is the name for the mineral whereas kaolin applies specifically to a rock made up of more than one mineral, or to the mineral group. To avoid confusion, the term "kandite" was introduced as the mineral group name, but it has not been universally accepted and consequently kaolin group is preferred.

Structure. The kaolinite unit layer consists of one silica tetrahedral sheet and one alumina octahedral sheet stacked such that apical oxygens from the tetrahedral sheet replace the hydroxyls in the octahedral sheet (**Fig. 1**). This stacked unit of one tetrahedral and one octahedral sheet is known as a 1:1 layer. The tetrahedral sheet is made up of individual silica tetrahedra linked to neighboring tetrahedra through three shared oxygens occupying a basal plane called the siloxane surface. The fourth, or apical, oxygen points upward in a direction normal to the siloxane surface and forms part of the adjacent octahedral sheet. The octahedral sheet is made up of octahedrally coordinated Al^{3+} ions linked together at the octahedral edges. In kaolinite there is little to no cation substitution and therefore essentially no layer charge. In the cases where substitution does occur, the most common are the substitution of Al^{3+} for Si^{4+} in the tetrahedral sheet and Fe^{2+} or Fe^{3+} for Al^{3+} in the octahedral sheet. Kaolinite exhibits varying amounts and types of stacking disorder related to translations between adjacent layers. Ordered kaolinite has very little stacking disorder compared to disordered kaolinite.

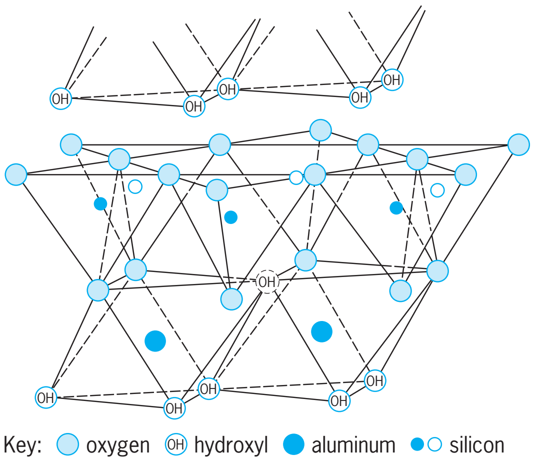


Fig. 1. Diagram of a kaolinite unit layer. (*After R. E. Grim, Applied Clay Mineralogy, McGraw-Hill, 1962*)

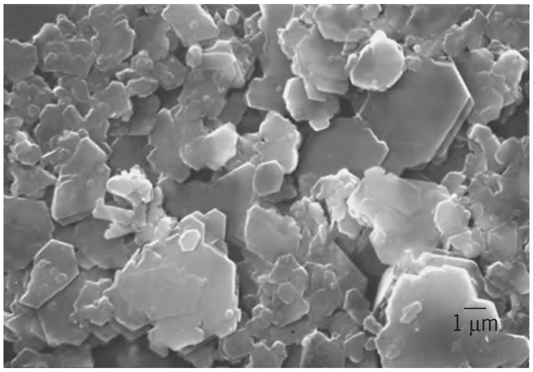


Fig. 2. Scanning electron micrograph showing pseudohexagonal kaolinite particles.

Properties. Kaolinite crystals are typically six-sided or pseudohexagonal in outline and may occur as single thin platelets or large accordionlike vermiform crystals (**Fig. 2**). Individual crystals range from approximately 0.1 to over 10 micrometers in diameter and from 15 to 50 nanometers in thickness. Kaolinite has a relatively high charge density and is therefore naturally hydrophilic, making it easily dispersible in water. It may be chemically modified so that it becomes hydrophobic and organophilic. It does not naturally swell in water unless treated with intercalation compounds such as hydrazine, urea, and ammonium acetate. Kaolinite is intrinsically white, has low plasticity, and has a refractive index of about 1.57. It has a low cation exchange capacity, and is chemically inert over a wide pH range except in some organic systems where it exhibits catalytic activity.

Genesis. Certain geochemical conditions favor the formation of kaolinite, which is most commonly found in surficial environments as a product of chemical weathering. Humid, temperate to tropical environments, where dilute, acid solutions are derived from the leaching of well-drained organic-rich soils, provide ideal conditions for kaolinite formation. The most common parent materials from which kaolinite is derived are feldspars and micas. These minerals, which contain the silica and alumina required for kaolinite formation, readily weather to kaolinite under the conditions listed above. Kaolinite also fills void spaces in sandstone, where it forms as a result of burial diagenesis. *See* DIAGENESIS.

Large accumulations of relatively pure kaolinite are classified as primary or secondary. Primary deposits are those derived from the in-situ alteration of feldspar-rich igneous, metamorphic, or sedimentary rocks such as granite, gneiss, or arkose. Alteration results from hydrothermal action, surface weathering, ground-water interaction, or a combination of these effects. Secondary deposits, also called sedimentary, are those that have been eroded, transported, and deposited as beds or lenses and are associated with sedimentary rocks. Primary deposits are more common than secondary deposits, which require special geologic conditions for their formation and preservation. For example, the extensive and high-quality paper-coating-grade deposits mined in the southeastern United States were most likely formed as the result of four events: (1) intense chemical weathering of the crystalline source area, (2) reworking and transportation of detrital grains by streams, (3) deposition of kaolinite-rich detritus along a passively subsiding margin, and (4) repeated cycles of oxidative weathering leading to recrystallization and continued kaolinization. *See* WEATHERING PROCESSES.

Uses. Kaolin is mined and beneficiated for a variety of industrial uses ranging from paper coating, to pharmaceuticals, to a source of aluminum for aluminum compounds. It is a versatile mineral that is also used in ceramics, plastics, paints, rubber, adhesives, caulks, catalysts, inks, gaskets, cosmetics, animal feeds, fertilizers, and fiberglass. This diverse range of applications is the result of a unique combination of factors, including particle size, particle shape, color, chemical composition, rheological properties, chemical inertness, thermal properties, and relatively low cost. Functionally kaolin plays many different roles in these products, including adding strength, imparting thixotropy, enhancing chemical resistance and electrical resistivity, and improving opacity. *See* ORE DRESSING; RHEOLOGY.

Because kaolin deposits naturally contain ancillary minerals along with kaolinite, at least some processing is required. Two processing options are available: wet and dry. The quality of the raw material and the desired properties of the final product dictate which process is used. Wet processing results in a purer, more uniform product and may employ a number of processing steps, including particle size separation, selective flocculation, magnetic separation, delamination, flotation, reductive leaching, calcination, and surface treatment. *See* CERAMICS; PARTICULATES.

Jessica Elzea Kogel

Bibliography. D. D. Carr (ed.), *Industrial Minerals and Rocks*, 6th ed., Society of Mining, Metallurgy and Exploration, 1994; R. E. Grim, *Applied Clay Mineralogy*, McGraw-Hill, 1962; D. M. Moore and R. C. Reynolds, *X-ray Diffraction and the Identification and Analysis of Clay Minerals*, Oxford University Press, 1997; H. H. Murray, W. Bundy, and C. Harvey (eds.), Kaolin Genesis and Utilization, *Clay Minerals Soc. Spec. Publ.*, no. 1, 1993; A. C. D. Newman (ed.), Chemistry of Clays and Clay Minerals, *Mineralog. Soc. Monog.*, no. 6, 1987.

Kapitza resistance

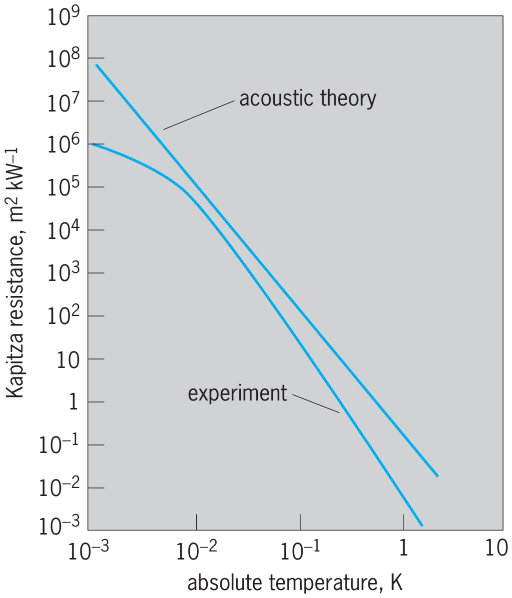
A resistance to the flow of heat across the interface between liquid helium and a solid. A temperature difference is required to drive heat from a solid into liquid helium, or vice versa; the temperature discontinuity occurs right at the interface. The Kapitza resistance, discovered by P. L. Kapitza, is defined in the equation below, where T_S and T_H are the solid

$$R_K = \frac{T_S - T_H}{\dot{Q}/A}$$

and helium temperatures and $\dot{Q}/A$ is the heat flow per unit area across the interface. *See* CONDUCTION (HEAT).

In principle, the measured Kapitza resistance should be easily understood. In liquid helium and solids (such as copper), heat is carried by phonons, which are thermal-equilibrium sound waves with frequencies in the gigahertz to terahertz region. The acoustic impedance of helium and solids can differ by up to 1000 times, which means that the phonons mostly reflect at the boundary, like an echo from a cliff face. This property together with the fact that the number of phonons dies away very rapidly at low temperatures means that at about 1 K there are few phonons to carry heat and even fewer get across the interface. The prediction is that the Kapitza resistance at the interface is comparable to the thermal resistance of a 10-m (30-ft) length of copper with the same cross section. *See* ACOUSTIC IMPEDANCE; PHONON; QUANTUM ACOUSTICS.

The reality is that above 0.1 K and below 0.01 K (10 mK) more heat is driven by a temperature difference than is predicted (see **illus.**). Above 0.1 K



Plot of the Kapitza resistance between copper and liquid helium-3 (^{3}He) as a function of absolute temperature. The acoustic theory is based upon heat transfer by thermal-equilibrium sound waves. The line labeled experiment is representative of many experimental results.

this is now understood to be a result of imperfections such as defects and impurities at the interface, which scatter the phonons and allow greater transmission. *See* CRYSTAL DEFECTS.

The enormous interest in ultralow-temperature (below 10 mK) research generated by the invention of the dilution refrigerator and the discovery of superfluidity in liquid helium-3 (^{3}He) below 0.9 mK also regenerated interest in Kapitza resistance, because heat exchange between liquid helium and solids was important for both the dilution refrigerator and superfluidity research. An ingenious technique was invented to overcome the enormous Kapitza resistance at 1 mK: The solid is powdered, and the powder is packed and sintered to a spongelike structure to enhance the surface area. In this way a 1-cm^3 (0.06-in.^{3}) chamber can contain up to 1 m^2 (10 ft^2) of interface area between the solid and the liquid helium.

It was found that at 1 mK the Kapitza resistance is 100 times smaller than predicted by the phonon model. There have been two explanations for the anomaly, and probably both are relevant. One is that energy is transferred by magnetic coupling between the magnetic ^{3}He atoms and magnetic impurities in the solid or at the surface of the solid; the other is that the spongelike structure has quite different, and many more, phonons than a bulk solid and that these can transfer heat directly to the ^{3}He atoms.

Whatever its cause, this anomaly has had a major impact on ultralow-temperature physics. On the one hand, it has allowed liquid ^{3}He to be cooled by adiabatic demagnetization refrigerators to well below 1 mK (the record is about 10 μK) so that its superfluid properties can be studied over a wide range of temperature. On the other, it has allowed the development of the dilution refrigerator so that it can operate to below 2 mK. In the dilution refrigerator the fluid that refrigerates is liquid ^{3}He. The purpose, however, is to refrigerate a sample or an experiment, and that requires heat exchange to the cold liquid ^{3}He. Such refrigerators are used to investigate the properties of matter at ultralow temperature. *See* ADIABATIC DEMAGNETIZATION; LIQUID HELIUM; LOW-TEMPERATURE PHYSICS; SUPERFLUIDITY.

John P. Harrison

Bibliography. D. F. Brewer (ed.), *Progress in Low Temperature Physics*, vol. 12, 1989; E. T. Swartz and R. O. Pohl, Thermal boundary resistance, *Rev. Mod. Phys.*, 61:605–668, 1989.

Kapok tree

Also called the silk-cotton tree (*Ceiba pentandra*), a member of the bombax family (Bombacaceae). The tree has a bizarre growth habit and produces pods containing seeds covered with silky hairs called silk cotton (see **illus.**). It occurs in the American tropics, and has been introduced into Java, Philippine Islands, and Sri Lanka. The silk cotton is the commercial kapok used for stuffing cushions, mattresses, and pillows. Kapok has a low specific gravity and is impervious to water, making it excellent for filling life preservers. *See* MALVALES.

Pods and leaves of kapok tree (*Ceiba pentandra*).

Perry D. Strausbaugh; Earl L. Core

Kármán vortex street

A double row of line vortices in a fluid. Under certain conditions, a Kármán street is shed in the wake of bluff cylindrical bodies when the relative fluid velocity is perpendicular to the generators of the cylinder (**Fig. 1**). This shedding of eddies occurs first from one side of the body and then from the other, an unusual phenomenon because the oncoming flow may be perfectly steady. Vortex streets can often be seen, for example, in rivers downstream of the columns supporting a bridge. The streets have been studied most completely for circular cylinders at low subsonic flow speeds. Regular, perfectly periodic vortex streets occur in the range of Reynolds number (Re) 50–300, based on cylinder diameter. Above a Reynolds number of 300, a degree of randomness begins to occur in the shedding due to secondary instabilities, which becomes progressively greater as Reynolds number increases, until finally the wake is completely turbulent. The highest Reynolds number at which some slight periodicity is still present in the turbulent wake is about 10^6. *See* REYNOLDS NUMBER.

Vortex streets and their effects are common. For example, they can be created by steady winds blowing past smokestacks, transmission lines, bridges, missiles about to be launched vertically, and pipelines aboveground in the desert. They are also related to the phenomenon that generates cloud

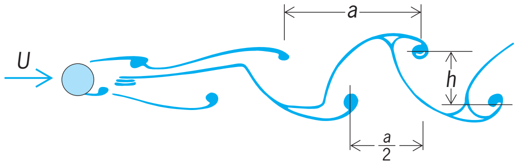


Fig. 1. Diagram of a Kármán vortex street. The streamwise spacing of the vortices, *h*, and the spacing normal to it, *a*, are shown.

Fig. 2. Vortex patterns downstream of Guadelupe Island, imaged by the *SeaWIFS* satellite.

patterns indicating the alternate shedding of vortices downstream of an island, as captured by satellite images (**Fig. 2**). Although the island only forms a surface obstruction to the wind, geophysical effects (Coriolis effects) generate a stagnant column of air in the atmosphere above it that creates an obstruction analogous to a cylinder. *See* CORIOLIS ACCELERATION.

The alternating vortex shedding that coincides with a well-formed vortex street produces oscillation in lateral forces acting on the body. If the vortex shedding frequency is near a natural vibration frequency of the body, the resonant response may cause large-amplitude vibration and in some cases structural damage. The Aeolian tones, or singing of wires in a wind, is an example of forced oscillation due to formation of a vortex street. Such forces can impose on structures unwanted vibrations, often leading to serious damage. An example of unwanted vibrations which led to serious damage is the collapse of the Tacoma Narrows (Washington) suspension bridge in 1940. This large structure made of thousands of tons of steel collapsed after a few hours of ever-growing oscillations induced by a steady wind (**Fig. 3**).

Fig. 3. Tacoma Narrows Bridge in the middle of its wild oscillations, a little before it collapsed.

T. von Kármán showed that an idealized, infinitely long vortex street is stable to small disturbances if the spacing of the vortices is such that $h/a = 0.281$; actual spacings are close to this value. A complete and satisfying explanation of the formation of vortex streets, however, has not yet been given. For $10^3 < \mathrm{Re} < 10^5$, the shedding frequency f for a circular cylinder in low subsonic speed flow is given closely by $fd/U = .21$, where d is the cylinder diameter and U is stream speed; a/d is approximately 5. This means that for a cylinder of a certain diameter immersed in a fluid, the frequency of vortex shedding is proportional to the velocity of the oncoming stream and inversely proportional to the diameter. A. Roshko discovered a spanwise periodicity of vortex shedding on a circular cylinder at $\mathrm{Re} = 80$ of about 18 diameters; thus, it appears that the line vortices are not quite parallel to the cylinder axis. *See* FLUID-FLOW PRINCIPLES; VORTEX.

Arthur E. Bryson, Jr.; Demetri P. Telionis

Bibliography. W. R. Debler, *Fluid Mechanics Fundamentals*, 1990; J. A. Liggett, *Fluid Mechanics*, 1994; R. L. Panton, *Incompressible Flow*, Wiley Interscience, 2d ed., 1996; F. M. White, *Fluid Mechanics*, McGraw-Hill, 6th ed., 2006.

Karst topography

Distinctive associations of third-order, erosional landforms indented into second-order structural forms such as plains and plateaus. They are produced by aqueous dissolution, either acting alone or in conjunction with (and as the trigger for) other erosion processes. Karst is largely restricted to the most soluble rocks, which are salt (360,000 mg/liter), gypsum and anhydrite (2400 mg/liter), and limestone and dolostone (30–400 mg/liter). The numbers denote effective solubility in meteoric waters under standard conditions. Being so soluble, salt is seen only in the driest places (for example, Death Valley, California), where it displays intensive dissolution topography. Surface gypsum karst is also comparatively rare in humid regions, but its dissolution at depth may produce widespread collapse and subsidence landforms (covered karst); it is often well developed in arid and semiarid areas such as the Pecos River valley of New Mexico. Pure limestones ($CaCO_3$) are the principal karst rocks, hosting the greatest extent and variety of features. Dolostone [the double carbonate mineral, $CaMg(CO_3)_2$] is significantly less soluble: the type and scale of dissolutional forms developed in it is normally less than in limestone, giving rise to topographies that are transitional to fluvial (stream-derived) landscapes. *See* DOLOMITE ROCK; GYPSUM; LIMESTONE.

Karst rocks outcrop over ~12% of the Earth's continental and island surfaces, but the distinctive topography is limited to about 8%. However, many well-known places are entirely karstic (the Bahamas and the majority of the Greek islands, for example), and 20–25% of the world's population relies on ground water circulating to some extent in karst rocks, so that karst studies are of considerable practical

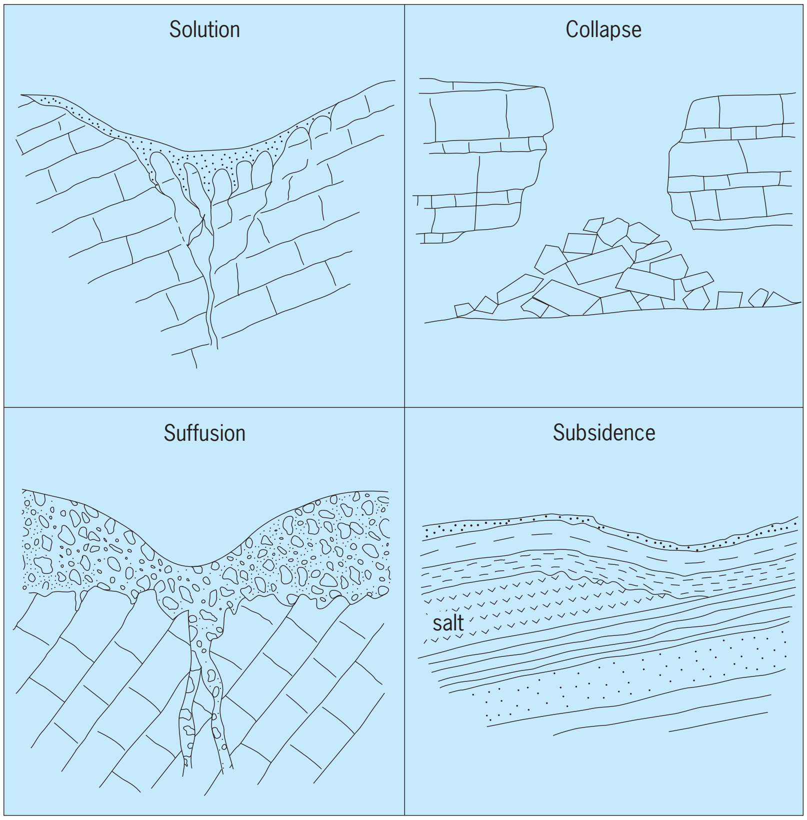


Fig. 1. **Principal genetic types of sinkholes. (*After D. C. Ford and P. W. Williams, Karst Geomorphology and Hydrology, 1989; copyright Chapman and Hall; used with kind permission from Kluwer Academic Publishers*)**

importance. *See* GROUND-WATER HYDROLOGY.

The essence of the karst dynamic system is that meteoric water (rain or snow) is routed underground, because the rocks are soluble, rather than flowing off in surface river channels. It follows that dissolutional caves develop in fracture systems, resurging as springs at the margins of the soluble rocks or in the lowest places. A consequence is that most karst topography is "swallowing topography," assemblages of landforms created to deliver meteoric water down to the caves. The word "karst" derives from a Slavic regional term, *kras*, meaning stony ground—stony because the former soil cover in western Slovenia and neighboring areas was lost into caves as a consequence of deforestation and overgrazing.

Karst landforms develop at small, intermediate, and large scales.

Karren. This is the general name given to small-scale forms—varieties of dissolutional pits, grooves, and runnels. Individuals are rarely greater than 10 m (30 ft) in length or depth, but assemblages of them can cover hundreds of square kilometers. On bare rock, karren display sharp edges; circular pits or runnels extending downslope predominate. Beneath soil, edges are rounded and forms more varied and intricate.

The largest karren are clefts following major joint sets. Where regular, shallow, and exposed, they create the appearance of a laid paving-limestone pavement, which is common in glaciated regions. Where deeper, they may become ruiniform (corridor karst) or, if the blocks between the clefts are tapered by runnelling dissolution, pinnacled. Dense clusters of pinnacles up to 20 m (65 ft) or more are the most spectacular karren, termed *shilin* (stone forests) in Yunnan, China, where the largest assemblages occur.

Particularly dense and complex pitting patterns occur along seashores. The finest development is on tropical coasts where formation of deep pits and sharp residual pinnacles is enhanced by biological action—phytokarst. Sharp dissolution notching may occur at the high-tide line.

Sinkholes. Also known as dolines or closed depressions, sinkholes are the diagnostic karst (and pseudokarst) landform. They range from shallow, bowllike forms, through steep-sided funnels, to vertical-walled cylinders (**Fig. 1**). Asymmetry is common. Individual sinkholes range from about 1 to 1000 m (3 to 3300 ft) in diameter and are up to 300 m (1000 ft) deep. Many may become partly or largely merged.

Sinkholes can develop entirely by dissolution acting downward into receptor caves, by collapse of cave roofs upward, by washing of unconsolidated fine sediments such as soil cover into underlying cavities, or by slow subsidence. Most karst sinkholes are produced by a combination of the first two or three processes.

On plains, sinkholes may be scattered like shell pits, or packed together with overlapping perimeters (polygonal karst, found in the Sinkhole Plain of Kentucky). Densities greater than 1000/km^2 are known. In mountainous country, deep polygonal karst can create a spectacular "egg carton" topography of steep-sided sinkholes (cockpits) between conical hills, termed *fengcong* (peak cluster) in southern China where the finest examples are found.

Dry valleys and gorges. These large-scale features are carved by normal rivers, but progressively lose their water underground (via sinkholes) as the floors become entrenched into karst strata. Many gradations exist, from valleys that dry up only during dry seasons (initial stage) to those that are without any surface channel flow even in the greatest flood periods (paleo-valleys). They are found in most plateau and mountain karst terrains and are greatest where river water can collect on insoluble rocks before penetrating the karst (allogenic rivers).

Poljes. A Serbo-Croatian term for a field, polje is the generic name adopted for the largest individual karst landform. Ideally, this is a topographically closed depression with a floor of alluvium masking an underlying limestone floor beveled flat by planar corrosion. Rivers may cross the alluvium, flowing from upstream springs to downstream sinks. The polje is flooded seasonally, when dissolutional undercutting takes place around the perimeter of enclosing limestone hills, and further alluvium is added to the insoluble shield protecting the floor. In this manner, the polje expands.

Ideal poljes up to a few kilometers in length are found in many karst regions. Chains of 10 km (6 mi) poljes and individuals up to 60 km (36 mi) in length occur in former Yugoslavia, where they are partly of tectonic origin—karstic adaptations

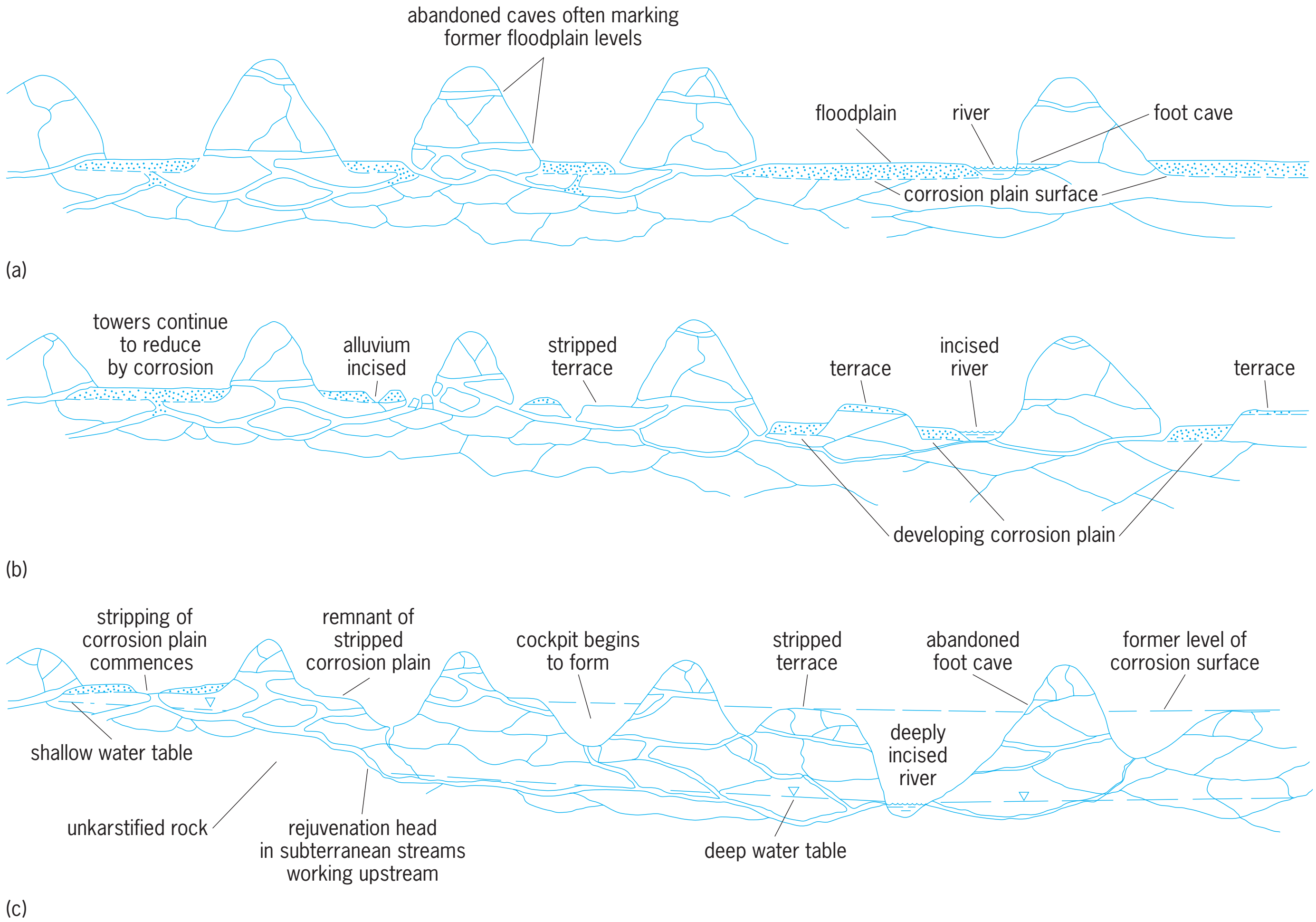


Fig. 2. Development of a multistage, polygenetic karst landscape of towers, cones, relict caves, corrosion plains, and terraces. (*After D. C. Ford and P. W. Williams, Karst Geomorphology and Hydrology, 1989; copyright Chapman and Hall; used with kind permission from Kluwer Academic Publishers*)

of downfaulted blocks. The Yugoslav examples are preeminent and are known as classical karst; they contain most of the cultivatable land in many areas, the intervening hills being unusable because of the great densities of karren and sinkholes. *See* FAULT AND FAULT STRUCTURES.

Karst plains and towers. These are the end stage of karst topographic development in some regions, produced by long-sustained dissolution or by tectonic lowering. The plains are of alluvium, with residual hills (unconsumed intersinkhole limestone) protruding through. Where strata are massively bedded and the hills are vigorously undercut by seasonal floods or allogenic rivers, they may be steepened into vertical towers. Tower karst (*fenglin* in Chinese) is the most accentuated karst topography, with individuals rising to as much as 500 m (1650 ft). The greatest tower karst development is in Guizhou and Guangxi provinces, China. Tower landscapes partly drowned by postglacial rising seas in Haiphong Bay, Vietnam, and Phuket, Thailand, are also famous tourist sights. Where strata are thinner and undercutting is ineffective, the hills tend to be conical. Plains with cone karst are common in Puerto Rico and Cuba.

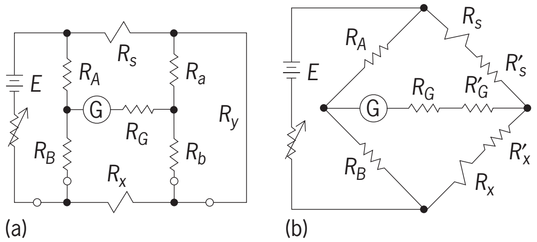
Tectonic uplift may induce incision by trunk rivers, producing complex multistage karst landscapes of towers, cones, and corrosion plains (**Fig. 2**). *See* CAVE; GEOMORPHOLOGY; WEATHERING PROCESSES.

Derek Ford

Bibliography. W. Dreybrodt, *Processes in Karst Systems*, Springer-Verlag, Berlin, 1988; D. C. Ford and P. W. Williams, *Karst Geomorphology and Hydrology*, Chapman and Hall, London, 1989; J. N. Jennings, *Karst Geomorphology*, Blackwell, Oxford, 1985; P. T. Milanovic, *Karst Hydrogeology*, Water Resources Publications, Littleton, CO, 1981; W. B. White, *Geomorphology and Hydrology of Karst Terrains*, Oxford University Press, 1988; D. Yuan, *Karst of China*, Geological Publishing House, Beijing, 1991.

Kelvin bridge

A specialized version of the Wheatstone bridge network designed to eliminate, or greatly reduce, the effect of lead and contact resistance and thus permit accurate measurement of low resistance. The circuit shown in **illus.** *a* accomplishes this by effectively placing relatively high-resistance-ratio arms in series with the potential leads and contacts of the low-resistance standards and the unknown



Kelvin bridge. (*a*) Actual circuit. (*b*) Equivalent Wheatstone bridge circuit.

resistance. In this circuit R_A and R_B are the main ratio resistors, R_a and R_b the auxiliary ratio, R_x the unknown, R_s the standard, and R_y a heavy copper yoke of low resistance connected between the unknown and standard resistors.

If the unwanted potential drop along the conductor linking R_s and R_x is divided (by R_a/R_b) in the same ratio (R_s/R_x) as that of the potential drops across R_s and R_x, the zero reading of the galvanometer (G) is unaffected by its actual value R_y.

Therefore, if R_a/R_b is adjustable (for example, by adjusting R_a) so that a null indication of G is obtained no matter what the value of R_y, the basic balance equation for the Wheatstone bridge (1) holds.

$$R_x = \frac{R_B}{R_A} R_s \tag{1}$$

A practical way of attaining this condition is to (1) adjust R_A/R_B with the lowest possible value of R_y (for example, a low-resistance copper link), and then (2) adjust R_a/R_b with R_y infinite (for example, with the link open-circuited). Steps 1 and 2 are carried out alternately as necessary until G is nulled for both conditions without significant further adjustments. The final adjustment of R_A/R_B obtains the bridge balance condition, Eq. (1).

A formal network analysis can be carried out by applying a delta-wye transformation to the network consisting of R_a, R_b, and R_y. The equivalent Wheatstone bridge network shown in illus. *b* is obtained, where Eqs. (2) hold. By an analysis similar to that for

$$\begin{aligned} R'_s &= \frac{R_y R_a}{R_y + R_a + R_b} \\ R'_x &= \frac{R_y R_b}{R_y + R_a + R_b} \\ R'_G &= \frac{R_a R_b}{R_y + R_a + R_b} \end{aligned} \tag{2}$$

the Wheatstone bridge, it can be shown that for a balanced bridge Eq. (3) holds. If Eq. (4) is valid, the

$$\begin{aligned} R_x = {} & \frac{R_B}{R_A} R_s \\ & + R_y \left(\frac{R_b}{R_a + R_b + R_y} \right) \left(\frac{R_B}{R_A} - \frac{R_b}{R_a} \right) \end{aligned} \tag{3}$$

$$\frac{R_B}{R_A} = \frac{R_b}{R_a} \tag{4}$$

second term of Eq. (3) is zero, the measurement is independent of R_y, and Eq. (1) is obtained.

As with the Wheatstone bridge, the Kelvin bridge for routine engineering measurements is constructed using both adjustable ratio arms and adjustable standards. However, the ratio is usually continuously adjustable over a short span, and the standard is adjustable in appropriate steps to cover the required range. *See* WHEATSTONE BRIDGE.

Sensitivity. The Kelvin bridge sensitivity can be calculated similarly to the Wheatstone bridge. The open-circuit, unbalance voltage appearing at the detector terminals may be expressed, to a close degree of approximation, as in Eq. (5).

$$e = E \frac{r}{(r+1)^2} \frac{\Delta R_x}{R_x + R_y \left(\dfrac{r}{r+1} \right)} \tag{5}$$

The unbalance detector current for a closed detector circuit may be expressed as in Eq. (6).

$$I_G = \frac{E \left(\dfrac{\Delta R_x}{R_x} \right)}{\dfrac{R_G}{r/(r+1)^2} + R_A + R_B + R_a + R_b} \tag{6}$$

The Kelvin bridge requires a power supply capable of delivering relatively large currents during the time a measurement is being made. The total voltage applied to the bridge is usually limited by the power dissipation capabilities of the standard and unknown resistors.

Errors. Kelvin bridge resistance-measurement errors are caused by the same factors as for the Wheatstone bridge. However, additional sources of error, as implied by the second term of Eq. (2), must be evaluated since these factors will seldom be reduced to zero. For minimum error the yoke resistance should be made as low as possible by physically placing the commonly connected current terminals of the unknown and standard as close together as possible and connecting with a low-resistance lead.

The ratio resistors each include not only the resistance of the resistors but also that of the interconnecting wiring and external leads and the contact resistance of the potential circuit contacts. The external leads are most likely to cause errors, and they should therefore be of the same resistance so that Eq. (3) will be fulfilled as nearly as possible. In addition, they should be relatively short, since the addition of a large resistance (long leads) will introduce an error in the calibrated ratio R_B/R_A. For precise measurements, trimmer adjustments are required in the ratio-arm circuits and provision is made to connect the bridge resistors into two different Wheatstone bridge configurations. By successively balancing first the Kelvin network and then each of the Wheatstone networks, these additive errors are virtually eliminated. *See* BRIDGE CIRCUIT; RESISTANCE MEASUREMENT. Charles E. Applegate; Bryan P. Kibble

Kelvin's circulation theorem

A theorem useful for ideal flows that have negligible viscous forces and a nearly constant density (incompressible flows). Such flows are potential flows and are applicable to the aerodynamics of wings. One characteristic of potential flows is a lack of vorticity. Vorticity characterizes the solid-body-like rotation of the fluid particles as they translate and deform. If a small element of fluid in a flow were instantaneously frozen, vorticity would indicate its spinning motion. Potential flows have zero vorticity. *See* POTENTIAL FLOW; VORTICITY; WING.

Circulation is a number associated with the spinning motion of a large region of a flow. Consider a plane flow such as the flow around a wing that spans a very big wind tunnel (**Fig. 1**). Imagine that you choose a certain closed path to walk around in a counterclockwise direction. As you walk at a uniform pace, measure and record the wind velocity component that is in the same direction as your path. Ignore the side component. Wind velocity in the same direction as your path is positive, while wind velocity against your path is negative. When you return to the starting point, the average velocity number indicates the circulation Γ for that circuit. Actually, the circulation Γ is the average wind component times the length of the path. Mathematically it can be shown that the circulation is the area integral of the vorticity over the region bounded by the closed-path curve, hence the interpretation as the global spinning character of a large region. In potential flows, the average velocity along the path is always zero for any path within the potential flow region. The average flow against the path is the same as the average flow with the path.

Kelvin's circulation theorem states that the circulation for a circuit of material particles is constant with time in an ideal flow. That is, the average rate of spinning of a piece of fluid is constant as it moves with the flow. This is a direct result of the lack of friction.

As an example, consider a plane vortex flow. This is a swirling potential flow of circular streamlines with zero vorticity everywhere except for the central core, where the vorticity is very high because of viscous forces. Any closed circuit that excludes the core has a circulation of zero. Any closed circuit that includes the core has the same finite circulation. The circulation is a measure of the strength of the vortex.

Consider a wing, helicopter blade, or fan blade starting from rest. The original and current positions are shown in **Fig. 2**. If a fluid is at rest, the circulation is zero for any circuit. A region of potential flow starting from rest will continue to have a circulation of zero for all time. As the wing starts moving, viscous forces are important in a boundary layer near the surface and produce a thin region of vorticity. The boundary layer fluid is not an ideal potential flow. This region is swept off and forms the core of a starting vortex that remains near the starting position. A thin viscous wake connects the starting vortex to the viscous boundary layer region that continues with the wing. A circuit around the wing—outside the boundary layer, the wake, and the starting vortex core—has zero circulation. However, a circuit cutting the viscous wake and including the vortex has a circulation of one sense; and a circuit cutting the wake and including the wing has a circulation of equal value but opposite sense. Kelvin's theorem does not apply to the vortex core, wake, or boundary layer particles because they have experienced viscous forces. *See* BOUNDARY LAYER FLOW.

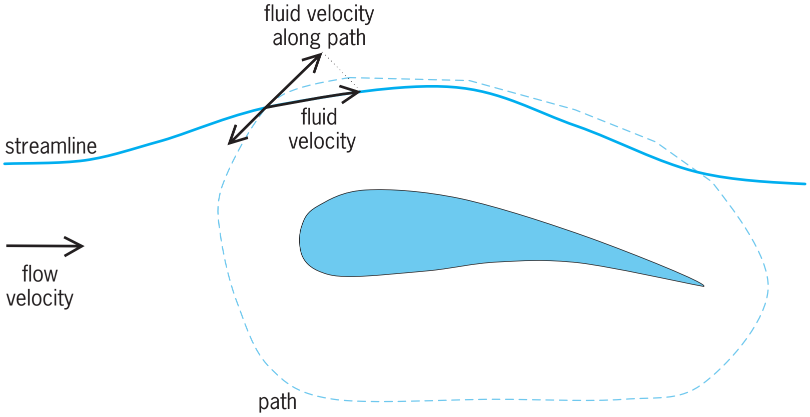


Fig. 1. Cross-section view of a wing in a wind tunnel depicting a circulation path.

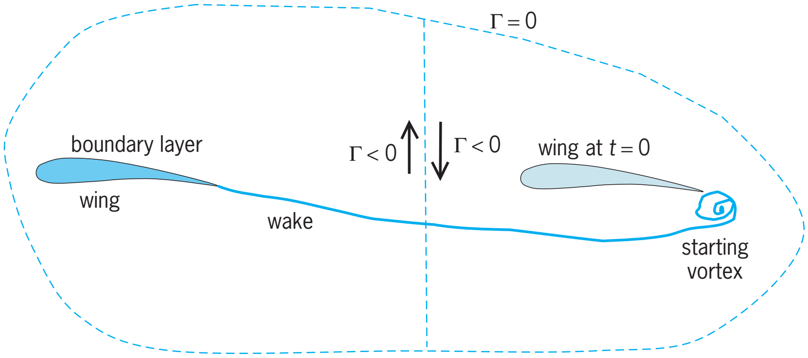


Fig. 2. Cross-section view of a wing starting from rest in a still fluid and leaving a starting vortex.

The nonzero circulation actually is directly related to the amount of lift produced by the wing. Circulation implies that the velocities above the wing are higher than those below the wing. The increase in velocity in potential flow is accompanied by a decrease in pressure. Thus, the lower pressures on the top result in a lift force. *See* BERNOULLI'S THEOREM.

A wing has a finite span, and so the disturbed flow does not extend to infinity in the spanwise direction (**Fig. 3**). A circuit path outside the wing/wake/

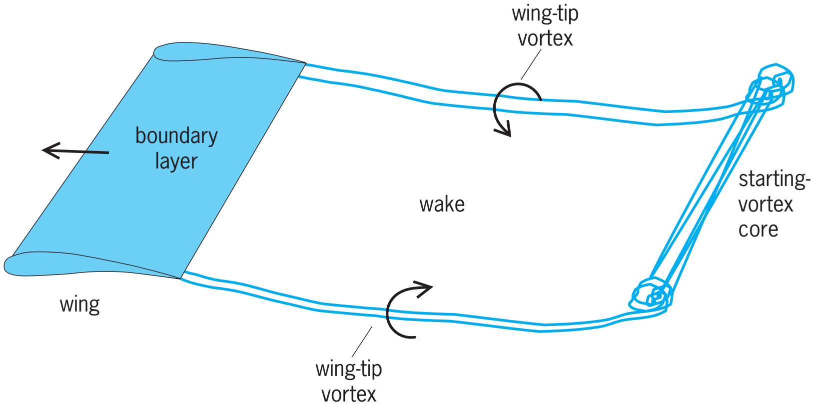


Fig. 3. Wing of finite span showing tip vortices connecting the wing circulation to the starting vortex.

starting vortex has zero circulation, because no viscous forces have touched those particles. However, the starting vortex and the wing are connected by a tip vortex so that a circuit around the tip vortex and through the wake has a finite circulation.

Ronald L. Panton

Bibliography. I. G. Currie, *Fundamental Mechanics of Fluids*, 3d ed., 2003; S. I. Green (ed.), *Fluid Vorticies*, 1995; H. J. Lugt, *Vortex Flows in Nature and Technology*, 1983; R. L. Panton, *Incompressible Flow*, 3d ed., 2005; P. G. Saffman, *Vortex Dynamics*, 1992.

Kelvin's minimum-energy theorem

A theorem in fluid dynamics that pertains to the kinetic energy of an ideal fluid (that is, inviscid, incompressible, and irrotational) and provides uniqueness statements concerning the solution of potential-flow problems.

Much of the value of potential theory lies in the wealth of theorems and corollaries which have been derived in the various fields by using this theory as their basis. Given that the velocity field can be represented by harmonic potential functions, Green's first identity leads to a result known as the energy theorem. This theorem states that the energy of the fluid in a region bounded by a surface *S* is equal to the product of half the fluid density and an integral over *S*; the integrand of this integral is the product of the velocity potential and its partial derivative in the direction of the surface normal. (Since the fluid velocity is the gradient of the velocity potential, the normal derivative of this potential equals the normal component of the fluid velocity at the boundary *S*, and therefore determines the normal motion of this boundary.) Using the energy theorem, Lord Kelvin derived a remarkable theorem which demonstrated that the irrotational motion of a liquid occupying a simply connected region has less kinetic energy than any other motion consistent with the same normal motion of the boundary *S*. *See* GREEN'S THEOREM; POTENTIALS.

The implications of the minimum-energy of irrotational motion are: (1) irrotational motion is impossible in a simply connected region bounded by fixed walls since in this case the normal derivative of the velocity potential vanishes at all points on the boundary, and therefore, according to the energy theorem, the kinetic energy is zero, or the system is at rest; (2) irrotational motion is impossible in a fluid in which the velocity at infinity vanishes if the internal boundaries are also at rest; (3) if the velocity at infinity vanishes, then the irrotational motion due to prescribed motion of an internal boundary is unique; and (4) if a fluid is in motion with uniform velocity at infinity, then the irrotational motion due to prescribed motion of an internal boundary is unique. *See* FLUID-FLOW PRINCIPLES; LAPLACE'S IRROTATIONAL MOTION.

Eric Paterson; Fred Stern

Bibliography. H. Lamb, *Hydrodynamics*, 6th ed., 1993; H. Rouse, *Advanced Mechanics of Fluids*, 1976.

Kenaf

An annual, short-day, herbaceous plant (*Hibiscus cannabinus*) of the Malvaceae family, cultivated for its stem fibers. The genus *Hibiscus* has approximately 200 species, of which one, roselle (*H. sabdariffa* var. *altissima*), also is occasionally referred to as kenaf.

Kenaf usually grows upright up to 15 ft (5 m) in height, and is cylindrical and either branched or unbranched. The stem is composed of two fibers, bast and core. The bast fibers, located in the bark, are long compared to the core fibers, produced in the stem's interior. The leaves are either entirely heart-shaped or display radiating lobes. Flowers are typically yellow with deep red centers. Wild forms of kenaf are found in east and central Africa, where for several centuries kenaf has been used for both fiber and food. Selection and breeding work have developed varieties with higher fiber yields, improved disease resistance, and reduced branching.

Commercial uses. Kenaf is grown commercially for fiber production in many areas of the world, with the largest producer being China. In the United States, kenaf production is located in Texas, Louisiana, Mississippi, and California. Additional uses of kenaf include the manufacturing of various paper and pulp products, and as poultry litter, potting soil amendments, chemical- and oil-spill absorbents, animal and horse bedding, and packing materials. Other potential uses include the manufacturing of filters and particle and insulation boards. Kenaf leaves, which contain 20–30% crude protein, also may have potential as a livestock feed source.

Cultivation. Kenaf can adapt to a wide range of climates and soils. However, because kenaf cannot tolerate frost, planting should not occur at temperatures below 32°F (0°C). Optimum yields of kenaf are generally obtained on well-drained soils with fertility levels necessary to meet the nutritional requirements. In addition, most varieties of kenaf are photoperiod-sensitive, and vegetative growth increases until the daylight period becomes less than 12 h 30 min. Flowering is then initiated and the vegetative growth rate declines. Early planting maximizes yields by increasing the growing season. Several photoperiod-insensitive varieties have been developed for high-latitude and equatorial areas. Kenaf is propagated by seed and must be replanted annually. Dense plant populations generally produce greater total and bast fiber yields, reduce weed populations, and improve harvesting efficiency.

Harvesting in the United States is accomplished with modified sugarcane or forage chopping equipment. After cutting and drying the stalks, they are processed at fiber separation mills to meet the specific markets. Since the 1970s, there has been an ongoing attempt to focus on kenaf as an alternative nonwood fiber source for various industrial applications.

Diseases. Kenaf serves as a host to several diseases and nematodes. The southern root-knot nematode (*Meloidogyne incognita*) is probably the greatest constraint to kenaf production throughout the world. Breeding efforts to identify a high level of resistance have not yet been successful. However, several strains which possess significant field tolerance to the root-knot nematode/soil-borne fungi complex have been identified. Crop rotations with nonhost species have provided the best method of reducing yield losses caused by both nematodes and soil-borne pathogens. Breeding and selection work has resulted in varieties and breeding strains with resistance to some of the major diseases of kenaf, including anthracnose (caused by *Collectotrichum gloesporiodes*) and charcoal rot (caused by *Macrophomina phaseolina*). *See* MALVALES; PLANT PATHOLOGY. Charles G. Cook

Bibliography. J. M. Dempsey, Kenaf, *Fiber Crops*, 1975; G. B. Killinger, Kenaf (*Hibiscus cannabinus* L.), A multi-use crop, *Agron. J.*, 61:734–736, 1969; C. S. Taylor and D. E. Kugler, Kenaf: Annual fiber crop products generate a growing response from industry, in *U.S. Department of Agriculture Yearbook*, pt. III: *New Crops, New Uses, New Markets*, 1992; U.S. Department of Agriculture, Agricultural Research Service, Kenaf paper: A forest saving alternative, *Agr. Res.*, 36:6–8, 1988; F. D. Wilson and M. Y. Menzel, Kenaf (*Hibiscus cannabinus*), roselle (*Hibiscus sabdariffa*), *Econ. Bot.*, 18:80–91, 1964.

Kennel cough

A common, highly contagious respiratory disease of dogs, also known as canine infectious tracheobronchitis. Several different bacteria and viruses are usually associated with the disease. Symptoms are generally mild but may vary widely depending on the agent, the host, and environmental factors. The main feature of the disease is sudden onset of violent coughing in dogs that had a recent exposure to other (infected) dogs. The disease is easily transmitted between dogs by droplets in the air or direct contact and often occurs as outbreaks or as a seasonal infection. Most dogs completely recover within 2 weeks; however, chronic and severe forms of the disease sometimes occur.

Infectious agent. Infectious agents commonly associated with the disease are the bacterium *Bordetella bronchiseptica* and canine parainfluenza virus. These agents are widespread in dogs; in fact, few dogs can escape exposure to these agents. *Bordetella bronchiseptica* can, in addition, be found in many other species of animal such as cats, pigs, rats, guinea pigs, rabbits, horses, and rarely humans. Each agent is capable of producing a mild form of the disease; however, most single-agent infections probably show no symptoms of disease. Several species of mycoplasmas have been isolated from the lower respiratory tract of dogs with kennel cough, but always in combination with another agent (for example, bordetella or canine parainfluenza virus). These mycoplasmas are normally found in the upper respiratory tract of healthy dogs. *See* BORDETELLA; MYCOPLASMAS.

Pathogenesis. Replication of canine parainfluenza virus, bordetella, and mycoplasma is usually limited to the epithelium lining the airway. Canine parainfluenza virus ultimately destroys the epithelial cells, while bordetella and mycoplasma remain attached to the outer surface of cells and cause more subtle changes in cell function such as reduced ciliary movement, attraction of inflammatory cells to the tissue, and alteration of phagocytic activities.

Epidemiology. Close contact with other dogs is usually required for transmission of kennel cough. Each of the viral agents of the disease, and possibly some of the mycoplasmas, have host ranges restricted to dogs. Because infections that show no symptoms of disease are also common, it is sometimes difficult to determine the source of the infection. Canine parainfluenza virus and *B. bronchiseptica* do not usually persist longer than a few weeks or a few months, respectively, in an individual dog. Dogs that have recovered from infection develop short-term immunity to reinfection. Therefore, infectious tracheobronchitis is acquired, and the pool of infectious agents is maintained in nature, by situations that bring infected dogs into close contact with susceptible dogs, for example, in closed breeding colonies, pet shops, boarding kennels, and show arenas.

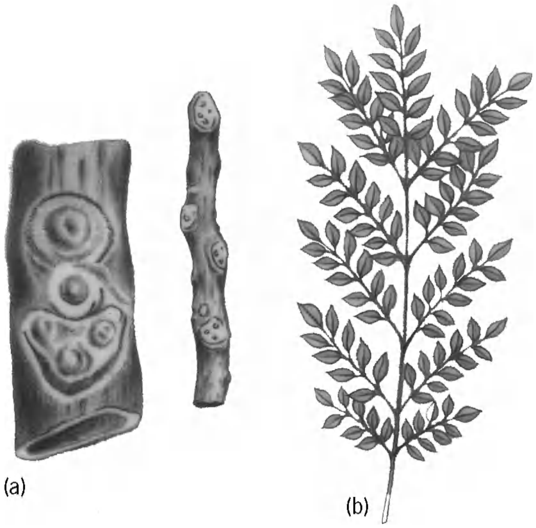
Diagnosis. The most common symptom of kennel cough is often described as a dry hacking or honking cough that may be accompanied by significant retching, as if a bone or other foreign object were lodged in the upper airway. Occasionally, there are discharges from the nose and eyes. Fever is not usually present, and dogs recover spontaneously in a few days to a few weeks. Kennel cough does not appear to have age, breed, geographic, or seasonal limitations; however, other factors, such as air quality, temperature change, stress, and other underlying diseases, may contribute to the onset and severity of disease. Under these circumstances, agents of the disease, particularly bordetella and mycoplasma, may be associated with protracted (chronic) cough or life-threatening pneumonia.

Treatment. Treatment of kennel cough is often unwarranted. However, when indicated, antitussives, bronchodilators, and corticosteroids are used to relieve coughing, and antimicrobials are used to treat or prevent bronchopneumonia. Apart from maintaining dogs in good general health, free of other underlying diseases, the risk of acquiring kennel cough can be reduced by minimizing exposure to infectious agents. This includes isolating infected dogs, disinfecting contaminated areas and objects, avoiding overcrowded conditions, increasing ventilation, and isolating dogs entering a new environment. Regular vaccinations for immunization against canine parainfluenza virus and *B. bronchiseptica* is also important in controlling the disease. David A. Bemis

Bibliography. C. E. Greene (ed.), *Infectious Diseases of the Dog and Cat*, 2d ed., 1998.

Kentucky coffee tree

A large, strikingly distinct tree, *Gymnocladus dioica*, which usually grows 80–90 ft (24–27 m) high, but sometimes attains a height of 110 ft (33 m) and a diameter of 5 ft (1.5 m). The species name, *dioica*, means that the tree is dioecious; that is, male and female flowers are on different individuals. It grows from eastern Nebraska, Kansas, and Oklahoma to southern Ontario, western New York, and Pennsylvania, and southwestward to Louisiana. It can readily be recognized when in fruit by its leguminous pods containing hard, heavy, red-brown seeds, which were used by early settlers as a substitute for coffee; hence the name coffee tree. The branches are stout and thick, and the bark has thin, twisted ridges. The leaves are twice pinnate, and the winter buds, sunken in the bark, are superposed, two or three together (see **illus.**). Never a common tree, it is sometimes cultivated in parks and gardens of the eastern United States and northern and central Europe. It is sometimes used as a street tree. The wood is hard and reddish, and is used for construction. It is durable in contact with the soil and is also used for railroad ties and fenceposts. *See* FOREST AND FORESTRY; ROSALES; TREE.

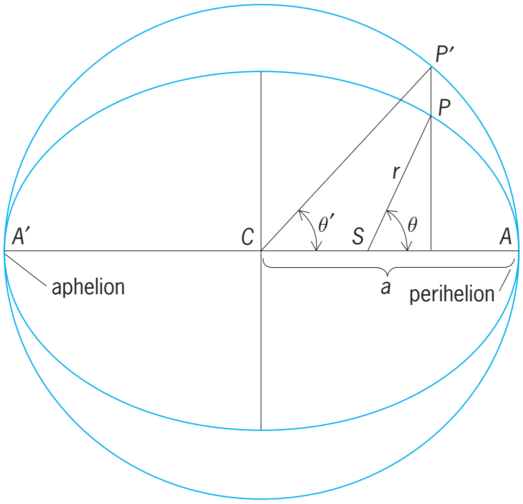


Kentucky coffee tree (*Gymnocladus dioica*). (*a*) Parts of branch showing leaf scars and winter buds. (*b*) Branch.

Arthur H. Graves; Kenneth P. Davis

Kepler's equation

The mathematical relationship between two different systems of angular measurement of the position of a body in an ellipse; specifically the relation between the mean anomaly M and eccentric anomaly θ', $M = \theta' - e \sin \theta'$, where e is the eccentricity of the ellipse. *See* ELLIPSE.



True and eccentric anomaly for a planet *P*.

The true position of a planet P in an elliptical orbit can be represented by the angle θ (true anomaly) measured at the focus S between the line directed to the planet and the line directed to the perihelion A (see **illus.**). The radius vector r from the focus to the planet can be expressed by $r = a(1 - e^2)/(1 + e \cos \theta)$, where a is the semimajor axis. The radius vector may also be expressed in terms of the eccentric anomaly θ' by $r = a(1 - e \cos \theta')$, where θ' is the angle at the center of the ellipse measured from perihelion along the circumscribed circle of radius a to the point whose projection perpendicular to the major axis passes through the planet. The true anomaly θ may be expressed in terms of the eccentric anomaly θ' by $\tan (\theta/2 = [(1 + e)/(1 - e)]_{1/2} \tan (\theta'/2)$. In actual practice, however, it is more convenient to describe the angular position of a planet in an elliptical orbit at any time t by means of its average angular velocity n (called mean motion) and the time T of last perihelion passage. This angle M (mean anomaly) is expressed by $M = n(t - T)$. Therefore, given the orbital elements a, e, n, and T, it is possible by means of Kepler's equation and the intermediary angle θ' to evaluate the true anomaly θ and the actual position of the planet in the orbit for any instant t. *See* PLANETARY PHYSICS.

Several practical methods for the iterative solution of this transcendental equation exist: (1) Starting with the value of M and denoting approximate values of θ' and $\sin \theta'$ by θ'_0 and $\sin M$, solve $\theta'_0 = M + e \sin M$. Denoting a second approximation to θ' by θ'_1, solve $\theta'_1 = M = e \sin \theta'_0$. Indicating the third approximation to θ' by θ'_2, solve $\theta'_2 = M + e \sin \theta'_1$, continuing the iteration until the required convergence of θ'_i is obtained. (2) Starting with M and an approximate value of θ', solve $M_0 = \theta'_0 - e \sin \theta'_0$, where M_0 corresponds to θ'_0. Set $M - M_0 = \Delta M_0$ and let $\Delta\theta'_0$ be a first correction to θ'_0, then $\Delta\theta'_0 = \Delta M_0/1 - e \cos \theta'_0$. Set $\theta'_1 = \Delta\theta'_0 + \Delta\theta'_0$, evaluate

M_1 corresponding to θ_1', form $M - M_1 = \Delta M_1$, and solve $\Delta\theta_1' = \Delta M_1/1 - e \cos \theta_1'$. Then $\theta_2' = \theta_1' + \Delta\theta_1'$, and the process is repeated until the required convergence of $\Delta\theta_i'$ is reached. *See* CELESTIAL MECHANICS.

Raynor L. Duncombe

Bibliography. P. Colwell, *Solving Kepler's Equation over Three Centuries*, 1993; J. M. A. Danby, *Fundamentals of Celestial Mechanics*, 2d ed., 1988.

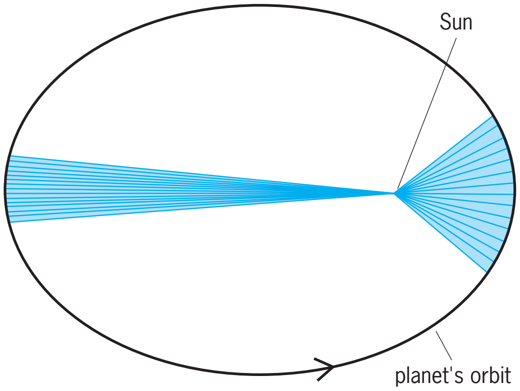
Kepler's laws

The three laws of planetary motion discovered by Johannes Kepler during the early years of the seventeenth century.

First law. The first law of Kepler states that a planet moves in an elliptical orbit around the Sun that is located at one of the two foci of the ellipse. An ellipse is one of the conic curves originally studied by Greek geometers. It is formed when a cone is cut by a plane that is neither parallel nor perpendicular to the axis and not parallel to the side of the cone. As the plane approaches perpendicularity with the axis, the ellipse approaches a circle; Kepler's first law accepts a circular orbit as the limiting case of an ellipse. As the cutting plane approaches a position parallel to the side of the cone, the ellipse approaches a parabola; although a body can move in a parabolic path around another placed at its focus, it would not be a planet in the accepted sense because it would never return. An ellipse is also defined as the locus of points the sum of whose distances from two fixed points (the foci) is a constant. *See* CONIC SECTION; EL-LIPSE.

Kepler discovered the first law by analyzing the observations that Tycho Brahe had made of the planet Mars in the late sixteenth century. An earlier astronomer of that century, Nicolaus Copernicus, had broken with the commonsense notion, apparently supported by daily experience, that the Earth is the center of the universe around which all other bodies move. However, he had accepted another tradition in astronomy, that all heavenly appearances are to be explained by combinations of circles, which were held to be perfect figures. Kepler's ellipses shattered that tradition. They also challenged another premise of earlier astronomy, that the heavens are fundamentally different from the Earth. Although Kepler discovered the ellipse by studying Mars, he generalized the shape to the orbits of all the planets.

In 1610, only one year after Kepler published his first law, Galileo discovered the satellites of Jupiter, and later in the seventeenth century satellites were observed around Saturn. All of these bodies obey Kepler's first law. In 1687, Isaac Newton demonstrated that any body, moving in an orbit around another body that attracts it with a force that varies inversely as the square of the distance between them, must move in a conic section. This path will be an ellipse when the velocity is below a certain limit in relation to the attracting force. Thus the first law is not limited to primary planets around the Sun or to heavenly bodies. It is a general law that applies to all satellites held in orbit by an inverse-square force.



Demonstration of Kepler's first and second laws. The planet moves along an elliptical orbit at a nonuniform rate, so that the radius vector drawn to the Sun, which is located at one focus of the ellipse, sweeps out areas that are proportional to time. Thus, the planet would take equal times (corresponding to the equal areas) to traverse the unequal distances along the ellipse that correspond to the two shaded areas. The diagram greatly exaggerates the eccentricity of any orbital ellipse in the solar system.

Second law. Astronomers needed to predict the location of planets at given times. The system of combining circles, which were assumed to turn uniformly, had made it possible to calculate a position at any time by the vectorial addition of the radii. In rejecting the system of circles, Kepler needed some other method by which to predict locations, and this he provided with his second law. It states that the radius vector of the ellipse (the imaginary line between the planet and the Sun) sweeps out areas that are proportional to time (see **illus.**). A planet does not move along its elliptical path at a uniform velocity; it moves more swiftly when it is closer to the Sun and more slowly when it is farther removed. Behind the nonuniform velocities the second law finds a uniformity in the areas described.

Again Newton demonstrated the dynamic cause behind Kepler's second law. In this case, it is not restricted to forces that vary inversely as the square of the distance; rather it is valid for all forces of attraction between the two bodies, regardless of the law the forces obey. The second law expresses the principle of the conservation of angular momentum. If body B moves in relation to body A in a straight line with a uniform velocity, it has a constant angular momentum in relation to A, and the line joining B to A sweeps out equal areas in equal increments of time. No force of attraction (or repulsion) between the two bodies can alter their angular momentum about each other. The second law then expresses a relation that holds for all pairs of bodies with radial forces between them. *See* ANGULAR MOMENTUM.

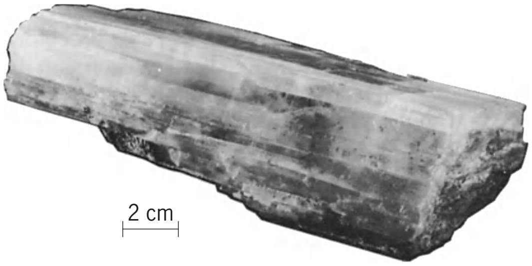
Third law. Kepler's first two laws govern the orbits of individual planets around the Sun. His third law defines the relations that hold within the system of planets. It states that the ratio between the square of a planet's period (the time required to complete one orbit) to the cube of the mean radius (the average distance from the Sun during one orbit) is a constant. The four satellites that Galileo had discovered around Jupiter were found to obey the third law, as did the satellites later found around Saturn. Newton demonstrated once again that the third law is valid for every system of satellites around a central body that attracts them, as the Sun attracts the planets, with a force that varies inversely as the square of the distance.

In fact, none of the three laws precisely describes the motions of the planets around the Sun, for the laws assume the only forces that act are those between the Sun and individual planets. But Newton's law of universal gravitation asserts that all planets also attract each other. In the case of the Moon orbiting the Earth, the attraction that the Sun exerts on the Moon is large enough to introduce perturbations, which astronomers had identified empirically before Kepler stated his three laws. Because the Sun is immensely more massive than the planets, the perturbations caused by the planets' mutual attractions are quite small, and Kepler's laws offer an accurate first approximation of their motions. Kepler's laws of planetary motion were the first of the mathematical laws of modern science; they did much to determine the pattern that science has continued to pursue. *See* CELESTIAL MECHANICS; GRAVITATION; ORBITAL MOTION; PERTURBATION (ASTRONOMY). Richard S. Westfall

Bibliography. G. Holton, Johannes Kepler's universe: Its physics and metaphysics, *Amer. J. Phys.*, 24:340–351, 1956; W. Pauli, *The Influence of Archetypal Ideas on the Scientific Theories of Kepler*, 1955; J. Voekel, *Johannes Kepler and the New Astronomy*, Oxford University Press, 2001.

Kernite

A hydrated borate mineral with chemical composition $Na_2B_4O_6(OH)_2{\cdot}3H_2O$. It is monoclinic with symmetry 2/m. It occurs only very rarely in crystals but is found most commonly in coarse, cleavable masses (see **illus.**) and aggregates. It has two perfect cleavages, {001} and {100}, at about 71° to each other, which together produce cleavage fragments that are elongated parallel to the *b* crystallographic axis. It is colorless to white; colorless and transparent specimens tend to become chalky white on exposure to air with the formation of tincalconite, $Na_2B_4O_5(OH)_4{\cdot}3H_2O$. The structure of kernite, with space group $P\,2/a$, contains infinite chains of composition $[B_4O_6(OH)_2]^{2-}$ which run parallel to the *b* crystallographic axis. The chains are linked to each other by bonding to Na^+ and by hydroxyl-oxygen bonds.



Typical cleavage fragment of kernite, from Boron, Kern County, California. (*Specimen from Department of Geology, Indiana University*)

Kernite was originally discovered at Boron, in the Kramer borate district, in the Mojave Desert of California. It is the second most important ore mineral, after borax, in this borate-producing region. It occurs with borax, $Na_2B_4O_5(OH)_4 \cdot 8H_2O$, and other borates such as ulexite and colemanite in a bedded series of Tertiary clays. Kernite occurs mainly near the bottom of the deposits and is believed to have formed by recrystallization from borax due to increased temperature and pressure. Kernite is also found in borate deposits in Tincalayu, in the province of Salta, Argentina. The three largest producers of crude borates are the United States, Turkey, and Argentina.

Boron compounds are used in the manufacture of glass, especially in glass wool used for insulation purposes. They are also used in soap, in porcelain enamels for coating metal surfaces, and in the preparation of fertilizers and herbicides. *See* BORATE MINERALS; BORON. Cornelis Klein

Bibliography. C. Klein and C. S. Hurlbut, Jr., *Manual of Mineralogy*, 21st ed., rev. 1999.

Kerogen

The complex, disseminated organic matter present in sedimentary rocks that remains undissolved by sequential treatment with common organic solvents (such as a benzene-methanol mixture, toluene, and methylene chloride) followed by treatment with nonoxidizing mineral hydrochloric acid and hydrofluoric acid. *See* SEDIMENTARY ROCKS.

The term kerogen is used to refer to a material derived from an extraction procedure. The term is defined operationally, and does not refer to a well-characterized chemical substance; thus some variability occurs in its definition. Some workers define kerogen as including all sediment organic matter that remains after extraction with organic solvents. However, kerogen isolation also involves recovery of the residual organic matter after treatment of the sediment with hydrochloric and hydrofluoric acids to remove the carbonate and silicate minerals, respectively. Therefore, the definition given here may be more appropriate in terms of the organic fraction actually isolated. The insoluble organic matter initially deposited with unconsolidated sediments often is not considered to be kerogen, but a so-called protokerogen that can be released by more drastic acid treatment (such as boiling several hours with 6 *N* hydrochloric acid) and is gradually converted to kerogen by elimination of carbon dioxide and water during early stages of sediment burial and diagenesis. However, the normal kerogen isolation procedures (using more dilute hydrochloric acid for shorter times) do not provide a clear distinction between kerogen and protokerogen, as defined above.

Kerogen is considered to be the major starting material for most oil and gas generation as sediments are subjected to geothermal heating in the subsurface. It is the most abundant form of organic carbon on Earth—about 1000 times more abundant than coal, which forms primarily from terrigenous remains of higher plants. Kerogen is formed from the remains of marine and lacustrine microorganisms, plants and animals, and variable amounts of terrigenous debris in sediments. The terrestrial portions of kerogen have elemental compositions similar to coal. *See* COAL.

Kerogen is sometimes pictured as a complex material of high molecular weight formed from the random condensation of monomers generated by an initial breakdown of polymeric biological precursor molecules following sediment burial. However, microscopic and chemical evidence is more in accord with a model of substantial incorporation of biological macromolecules that have undergone varying degrees of chemical alteration prior to and after burial.

Kerogens are classified according to their atomic ratios of hydrogen to carbon (H/C) and oxygen to carbon (O/C), with oil-prone kerogens being generally higher in H/C and lower in O/C than the gas-prone kerogens. With increasing length of exposure to subsurface temperatures, all kerogens show decreases in the O/C and H/C ratios as they generate preferentially carbon dioxide and water, then oil, and finally only gas (methane) at progressively higher subsurface depths and temperatures. *See* NATURAL GAS; PETROLEUM.

Jean K. Whelan

Kerosine

A refined petroleum fraction used as a fuel for heating and cooking, jet engines, lamps, and weed burning, and as a base for insecticides. Kerosine, known also as lamp oil, is recovered from crude oil by distillation. It boils in the approximate range of 350–550°F (180–290°C). Most marketed grades, however, have narrower boiling ranges. The specific gravity is about 0.8. Determined by the Abel tester, the flash point is not below 73°F (23°C), but usually a higher flash point is specified. Down to a temperature of −25°F (−32°C), kerosine remains in the liquid phase. Components are mainly paraffinic and naphthenic hydrocarbons which are in the C_{10}–C_{14} range. A low content of aromatics is desirable except when kerosine is used as tractor fuel.

Specifications are established for specific grades of kerosine by government agencies and by refiners. Since these specifications are developed from performance observations, they are adhered to rigidly to assure satisfactory operation. For use in lamps, for example, a highly paraffinic oil is desired because aromatics and naphthenes give a smoky flame; and for satisfactory wick feeding, a viscosity no greater than 2 centipoises is required in this application. Furthermore, the nonvolatile components must be kept low. In order to avoid atmospheric pollution, sulfur content must be low; a minimum flash point of 100°F (38°C) is desirable to reduce explosion hazards. *See* PETROLEUM PRODUCTS.

Harold C. Ries

Kerr effect

Electrically induced birefringence that is proportional to the square of the electric field. When a substance (especially a liquid or a gas) is placed in an electric field, its molecules may become partly oriented. This renders the substance anisotropic and gives it birefringence, that is, the ability to refract light differently in two directions. This effect, which was discovered in 1875 by John Kerr, is called the electrooptical Kerr effect, or simply the Kerr effect. *See* BIREFRINGENCE.

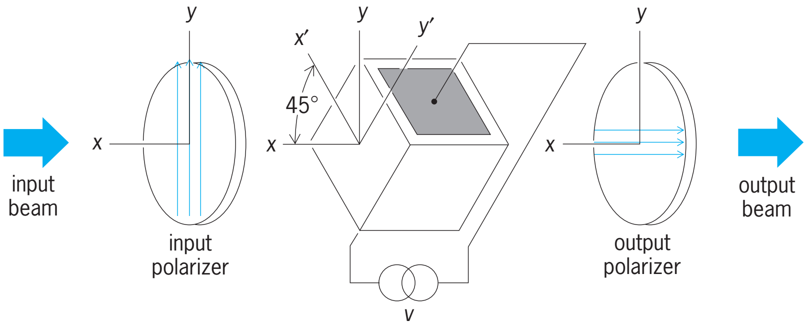
When a liquid is placed in an electric field, it behaves optically like a uniaxial crystal with the optical axis parallel to the electric lines of force. The Kerr effect is usually observed by passing light between two capacitor plates inserted in a glass cell containing the liquid. Such a device is known as a Kerr cell. There are two principal indices of refraction, n_o and n_e (known as the ordinary and extraordinary indices), and the substance is called a positively or negatively birefringent substance, depending on whether $n_e - n_o$ is positive or negative.

Light passing through the medium normal to the electric lines of force (that is, parallel to the capacitor plates) is split into two linearly polarized waves traveling with the velocities c/n_o and c/n_e, respectively, where c is the velocity of light, and with the electric vector vibrating perpendicular and parallel to the lines of force.

The difference in propagation velocity causes a phase difference δ between the two waves, which, for monochromatic light of wavelength λ_0, is $\delta = (n_e - n_o)x/\lambda_0$, where x is the length of the light path in the medium.

Kerr constant. Kerr found empirically that $(n_e - n_o) = \lambda_0 BE^2$, where E is the electric field strength and B a constant characteristic of the material, called the Kerr constant. Havelock's law states that $B\lambda n/(n-1)^2 = k$, where n is the refractive index of the substance in the absence of the field and k is a constant characteristic of the substance but independent of the wavelength λ. Roughly speaking, the Kerr constant is inversely proportional to the absolute temperature. The phase difference δ is determined experimentally by standard optical techniques. If the wavelength λ_0 is expressed in meters, and the field strength E in volts/m, the Kerr constant for carbon disulfide, which has been determined most accurately, is $B = 3.58 \times 10^{-14}$. Values of B range from -25×10^{-14} for paraldehyde to $+384 \times 10^{-14}$ for nitrobenzol.

The theory of the Kerr effect is based on the fact that individual molecules are not electrically isotropic but have permanent or induced electric dipoles. The electric field tends to orient these dipoles, while the normal agitation tends to destroy the orientation. The balance that is struck depends on the size of the dipole moment, the magnitude of



Kerr shutter. The arrows designate a light beam propagating through the cell. Polarizer axes are parallel to *x* and *y* axes; electric field of cell is parallel to *x*′ or *y*′ axis. (*After A. Yariv, Quantum Electronics, John Wiley and Sons, 1967*)

the electric field, and the temperature. This theory accounts well for the observed properties of the Kerr effect.

In certain crystals there may be an electrically induced birefringence that is proportional to the first power of the electric field. This is called the Pockels effect. In these crystals the Pockels effect usually overshadows the Kerr effect, which is nonetheless present. In crystals of cubic symmetry and in isotropic solids (such as glass) only the Kerr effect is present. In these substances the electrically induced birefringence (Kerr effect) must be carefully distinguished from that due to mechanical strains induced by the same field. *See* ELECTROOPTICS.

Kerr shutter. An optical Kerr shutter or Kerr cell consists of a cell containing a liquid (for example, nitrobenzene) placed between crossed polarizers. As such, its construction very much resembles that of a Pockels cell. With a Kerr cell (see **illus.**), an electric field is applied by means of an electronic driver v and capacitorlike electrodes in contact with the liquid; the field is perpendicular to the axis of light propagation and at 45° to the axis of either polarizer. In the absence of a field, the optical path through the crossed polarizers is opaque. When a field is applied, the liquid becomes birefringent, opening a path through the crossed polarizers. In commercial Kerr cell shutters, the electric field (a typical value is 10 kV/cm) is turned on and off in a matter of several nanoseconds (1 ns $= 10^{-9}$ s). For laser-beam modulation the Pockels cell is preferred because it requires smaller voltage pulses. Kerr cell shutters have the advantage over the Pockels cell of a wider acceptance angle for the incoming light. *See* OPTICAL MODULATORS.

A so-called ac Kerr effect or optical Kerr effect has also been observed and put to use in connection with lasers. When a powerful plane-polarized laser beam propagates through a liquid, it induces a birefringence through a mechanism that is very similar to that of the ordinary, or dc, Kerr effect. In this case, it is the ac electric field of the laser beam (oscillating at a frequency of several hundred terahertz) which lines up the molecules. By using laser pulses with durations of only a few picoseconds (1 ps $= 10^{-12}$ s), extremely fast optical Kerr shutters have been built in the laboratory. *See* LASER. Michel A. Duguay

Bibliography. A. Yariv, *Optical Electronics*, 5th ed., 1997.

Ketene

A member of a class of organic compounds with the C=C=O group as a common structural element. Ketenes are derivatives of carboxylic acids, from which they are (hypothetically) formed by abstraction of water; they can therefore be considered to be inner anhydrides of acids, as opposed to the common carboxylic acid anhydrides formed from two molecules of a carboxylic acid. An example is the formation of ketene and acetic anhydride from acetic acid [reactions (1) and (2)].

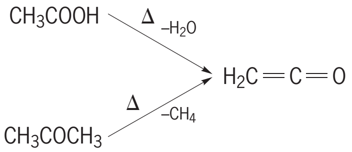
$$\mathrm{CH_3COOH} \longrightarrow \mathrm{H_2C{=}C{=}O} + \mathrm{H_2O} \qquad (1)$$

$$\mathrm{2CH_3COOH} \longrightarrow \mathrm{CH_3COOCOCH_3} + \mathrm{H_2O} \qquad (2)$$

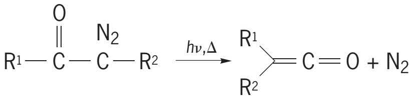
Infrared spectra of ketenes show very strong bands near 1120 and 2150 cm^{-1}, associated with the asymmetric stretching mode of the C=C and C=O bonds of the heterocumulene (C=C=O); both bonds are used for identification of ketenes. *See* INFRARED SPECTROSCOPY.

Like real anhydrides, ketenes are acylating agents which readily undergo reactions with many compounds containing active hydrogens. They are relatively labile compounds, and only a limited number have been prepared and isolated. Many ketenes have been prepared in place and reacted immediately. *See* ACID ANHYDRIDE.

Preparation. Ketene, the parent compound, is a colorless gas that boils at −41°C (−41.8°F) and is stable for an extent of time only at low temperatures. It is a commercial intermediate for the synthesis of acetic anhydride. Industrially it is produced by dehydration of acetic acid at 700–750°C (1290–1380°F) in the presence of triethyl phosphate as catalyst (90% selectivity). It can also be obtained by thermolysis of acetone at 600–700°C (1110–1290°F) in the presence of catalytic amounts of carbon disulfide [reaction (3)].

 (3)

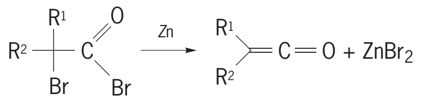
Other major laboratory methods, primarily for preparing the higher homologs, involve the thermolysis or radiative rearrangement of diazo carbonyl compounds (Wolff rearrangement) with loss of nitrogen [reaction (4)]; dehalogenation of 2-haloacyl

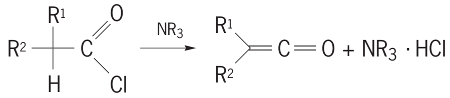
 (4)

$$\mathrm{R^1{-}\overset{\overset{\displaystyle O}{\|}}{C}{-}\overset{N_2}{C}{-}R^2} \xrightarrow{h\nu,\Delta} \mathrm{\begin{matrix}R^1\\ \\R^2\end{matrix}\!\!>\!C{=}O} + \mathrm{N_2} \qquad (4)$$

halides, predominantly bromides, with zinc [Zn;

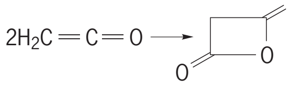
reaction (5)]; dehydrohalogenation of acyl halides with tertiary amines [NR_3; reaction (6)].


(5)

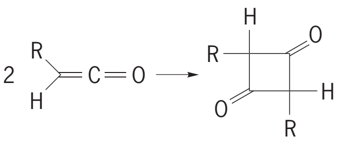

(6)

A variety of ketenes have been synthesized in which R^1 and R^2 in the formulas above represent hydrogen, alkyl, aryl, halogen, alkoxy, amino, carbonyl, thiocarbonyl, imino, cyano, or azido groups. Numerous ketenes in which the heterocumulene group is linked directly to metals such as silicon (Si), germanium (Ge), or tin (Sn) have also been prepared. Many alkyl ketenes have been prepared only in place and converted immediately into stable addition products. Aryl ketenes are more stable and can be prepared and isolated in relatively good yields. Haloketenes are extremely unstable and have most often been identified by their conversion into addition products. Nevertheless, dichloroketene has found wide application in the synthesis of various cycloadducts.

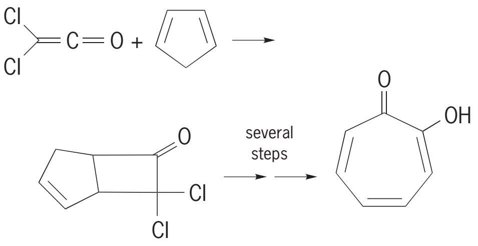
Reactions. Ketenes react readily with nucleophiles such as alcohols to form esters, with amines to form amides, and with acids to yield anhydrides. Reactions of ketenes that are important in chemical synthesis are 2 + 2 cycloadditions with a variety of double-bond systems. Four-membered-ring cycloadducts have been obtained with C═C, C═O, C═N, N═N, and N═S bonds. Primary examples are the dimerization of ketene, which produces an asymmetric adduct having the oxetanone structure [reaction (7)].


(7)

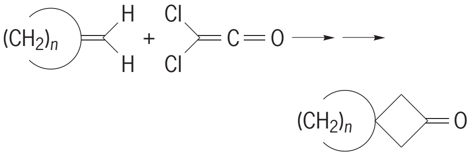
Mono- and disubstituted ketenes were found to form either symmetric dimers—cyclobutane derivatives, in which cycloaddition takes place only across the C═C double bond of the heterocumulene [reaction (8)]—or asymmetric dimers, or mixtures of both.


(8)

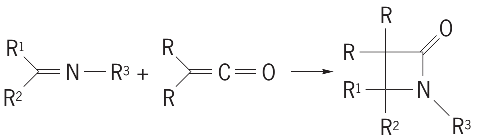
Halogenated ketenes do not dimerize but undergo cycloadditions with many other double-bond systems. These adducts are very valuable synthetic intermediates, since the halogens can be easily replaced in hydrogenation or substitution reactions. An example is the synthesis of tropolones (I) on the basis of the reaction of dichloroketene with cyclopentadiene [reaction (9)].


(9)

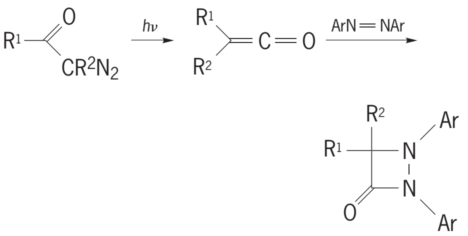
Spiro compounds have been prepared by reacting methylene cycloalkanes with dichloroketene [reaction (10)].


(10)

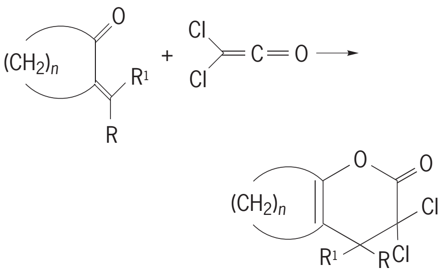
A large number of azetidinones (β-lactams) have been prepared via the 2 + 2 cycloaddition between ketenes and amidines and azomethines [reaction (11)].


(11)

Similarly, arylketenes, generated from diazoketones as shown in reaction (4), and diarylketenes react with azobenzene to yield a variety of 1,2,3-triaryl or tetraryl-1,2-diazetidinones [reaction (12)].


(12)

Ketenes also undergo 1,4-cycloaddition reactions with conjugated double-bond systems, which are used very often for the synthesis of both fused cyclic and heterocyclic systems [reaction (13)].


(13)

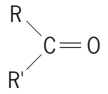
See STEREOCHEMISTRY. Reinhard H. Richter

Bibliography. S. Patai (ed.), *The Chemistry of Ketenes, Allenes, and Related Compounds*, pt. 1, 1980; M. B. Smith and J. March, *March's Advanced Organic Chemistry: Reactions, Mechanisms, and*

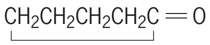
Structure, 5th ed., 2001; T. T. Tidwell, *Ketenes*, 2d ed., 2006.

Ketone

One of a class of chemical compounds of the general formula



R and R′ are alkyl, aryl, or heterocyclic radicals. The groups R and R′ may be the same or different or incorporated into a ring as in cyclopentanone (**1**).

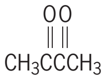


(**1**)

The ketones, acetone and methyl ethyl ketone, are used as solvents. Ketones are important intermediates in the syntheses of organic compounds.

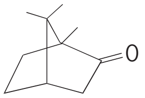
By common nomenclature rules, the R and R′ groups are named, followed by the word ketone, for example, $CH_3CH_2COCH_2CH_3$ (diethyl ketone), $CH_3COCH(CH_3)_2$ (methyl isopropyl ketone), and $C_6H_5COC_6H_5$ (diphenyl ketone). The nomenclature of the International Union of Pure and Applied Chemistry uses the hydrocarbon name corresponding to the maximum number of carbon atoms in a continuous chain in the ketone molecule, followed by "-one," and preceded by a number designating the position of the carbonyl group in the carbon chain. The first two ketones above are named 3-pentanone and 3-methyl-3-butanone.

Properties. The lower-molecular-weight ketones are colorless liquids. Acetone and methyl ethyl ketone are miscible with water; the water solubility of the higher homologs decreases with increasing number of carbon atoms. In the infrared absorption spectrum, the intense carbonyl stretching band near 1715 cm^{-1} is useful for identification and characterization of ketones. Because of their characteristic odors, various ketones are of use in the flavoring and perfumery industry. Biacetyl (**2**) is a principal ingredient in the flavoring of margarine, camphor (**3**) is valued for its medicinal odor, although it appears to have no therapeutic value, and cyclopentadecanone (Exaltone; **4**) has a musk odor and is used in perfumes.

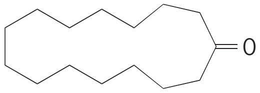


$$\mathrm{CH_3\overset{\overset{O}{\|}}{C}\overset{\overset{O}{\|}}{C}CH_3}$$

(**2**)



(**3**)

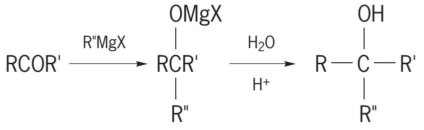


(**4**)

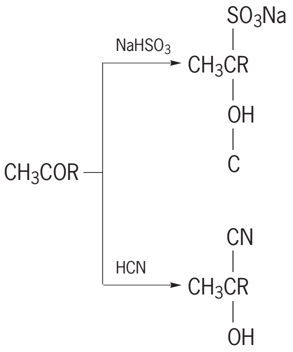
Reactions of ketones. Addition to the carbonyl group is the most important type of ketone reaction. Ketones are generally less reactive than aldehydes in addition reactions. Methyl ketones are more reactive than the higher ketones because of steric group effects.

Hydrogen adds catalytically to the carbonyl group, and lithium aluminum hydride gives the same type of product—the secondary alcohol, as shown in reaction (1). The Grignard reagent adds to the carbonyl group, and tertiary alcohols are formed by hydrolysis, reaction (2). Hydrogen cyanide and sodium bisulfite add to methyl ketones, as in reaction (3).

$$\mathrm{RCOR'} \xrightarrow[\text{or LiAlH}_4]{\mathrm{H_2(Ni\ or\ Pt)}} \mathrm{RCHOHR'} \qquad (1)$$



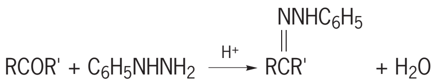
$$\mathrm{RCOR'} \xrightarrow{\mathrm{R''MgX}} \mathrm{R\underset{\underset{R''}{|}}{\overset{\overset{OMgX}{|}}{C}}R'} \xrightarrow[\mathrm{H^+}]{\mathrm{H_2O}} \mathrm{R{-}\underset{\underset{R''}{|}}{\overset{\overset{OH}{|}}{C}}{-}R'} \qquad (2)$$



(3)

Alcohols do not add readily to the ketone carbonyl as they do to the aldehyde carbonyl, but ketals may be formed by the action of orthoformates, as in reaction (4). Amine derivatives such as hydroxylamine (NH_2OH), phenylhydrazine ($C_6H_5NHNH_2$), and semicarbazide (NH_2-$CONHNH_2$) add to the carbonyl by breaking an N-H bond and with subsequent loss of water, as in reaction (5). The resulting oximes, phenylhydrazones, and semicarbazones are useful derivatives for the identification and characterization of ketones.

$$\mathrm{RCOR'} + \mathrm{HC(OC_2H_5)_3} \xrightarrow{\mathrm{H^+}} \mathrm{R{-}C(OC_2H_5)_2R'} + \mathrm{HCOOC_2H_5} \qquad (4)$$

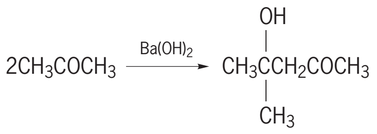


$$\mathrm{RCOR'} + \mathrm{C_6H_5NHNH_2} \xrightarrow{\mathrm{H^+}} \mathrm{R\overset{\overset{NNHC_6H_5}{\|}}{C}R'} + \mathrm{H_2O} \qquad (5)$$

Ketones supply alpha hydrogen in aldol-type condensation reactions, as in reaction (6), but can supply a carbonyl group only to a limited extent because

$$\mathrm{H_2CO} + \mathrm{CH_3COR} \xrightarrow{\mathrm{OH^-}} \mathrm{RCOCH_2OH} \xrightarrow{\mathrm{H^+}} \mathrm{RCOCH{=}CH_2} + \mathrm{H_2O} \qquad (6)$$

of its lower reactivity. An exception is the self-condensation of acetone to diacetone alcohol represented by reaction (7).



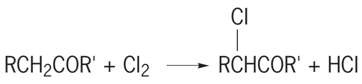
$$2CH_3COCH_3 \xrightarrow{Ba(OH)_2} CH_3\overset{OH}{\underset{CH_3}{C}}CH_2COCH_3 \qquad (7)$$

Ketones are oxidized less readily and less selectively than aldehydes, to give oxidation products such as carboxylic acids, by cleavage of the bonds from the carbonyl carbon atom to the adjacent atom.

Methyl ketones give the haloform reaction with solutions of iodine in aqueous potassium hydroxide, reaction (8).

$$CH_3COR \xrightarrow[KOH]{I_2} CHI_3 + RCOOK \qquad (8)$$

Chlorine or bromine will substitute the alpha hydrogen atoms of ketones, as in reaction (9). The re-



$$RCH_2COR' + Cl_2 \longrightarrow R\overset{Cl}{C}HCOR' + HCl \qquad (9)$$

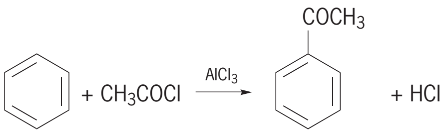
sulting α-haloketones are quite reactive in displacement reactions of the halogen atom.

Preparation. Ketones are formed by the dehydrogenation or oxidation of secondary alcohols, as shown in reaction (10), and this is the method

$$RCH(OH)R' \longrightarrow RCOR' \qquad (10)$$

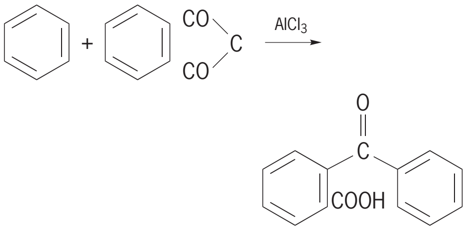
used industrially for the preparation of acetone and methyl ethyl ketone. In the United States, acetone and methyl ethyl ketone are produced primarily by this reaction.

Aromatic ketones may be prepared by the Friedel-Crafts acylation reaction using either acid halides as shown in reaction (11) for the preparation of ace-



(11)

tophenone, or anhydrides as shown in reaction (12) for the preparation of *o*-benzoylbenzoic acid.



(12)

Organocadmium reagents give ketones on reaction with acid halides, reaction (13).

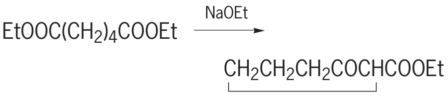
$$RCOCl + R'CdX \longrightarrow RCOR' + CdXCl \qquad (13)$$

Beta-ketoesters formed in the Claisen condensation are cleaved by aqueous sodium hydroxide solution to form ketones, reactions (14) and (15). The

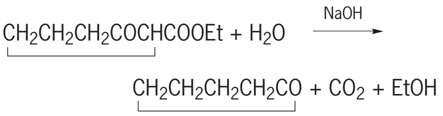
$$2RCH_2COOEt \xrightarrow{EtONa} RCH_2COCHRCOOEt + EtOH \qquad (14)$$

$$RCH_2COCHRCOOEt + H_2O \xrightarrow{NaOH} RCH_2COCH_2R + CO_2 + EtOH \qquad (15)$$

Dieckmann condensation of esters of dibasic acids leads in similar fashion to cyclic ketones, reactions (16) and (17).

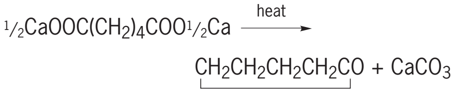


$$EtOOC(CH_2)_4COOEt \xrightarrow{NaOEt} \underbracket{CH_2CH_2CH_2COCH}COOEt \qquad (16)$$



$$\underbracket{CH_2CH_2CH_2COCH}COOEt + H_2O \xrightarrow{NaOH} \underbracket{CH_2CH_2CH_2CH_2CO} + CO_2 + EtOH \qquad (17)$$

Both cyclic and open chain ketones are formed by pyrolysis of calcium or thorium salts of dibasic acids or monobasic acids, reactions (18) and (19).



$${}^1/_2CaOOC(CH_2)_4COO{}^1/_2Ca \xrightarrow{heat} \underbracket{CH_2CH_2CH_2CH_2CO} + CaCO_3 \qquad (18)$$

$$2RCOO^1/_2Ca \xrightarrow{heat} RCOR + CaCO_3 \qquad (19)$$

See ALDEHYDE; FRIEDEL-CRAFTS REACTION; GRIGNARD REACTION; STERIC EFFECT (CHEMISTRY).

Paul E. Fanta

Bibliography. R. J. Fessenden and J. S. Fessenden, *Organic Chemistry*, 5th ed., 1994; R. T. Morrison and R. N. Boyd, *Organic Chemistry*, 7th ed., 2001.

Key telephone system

A communications system that allows users to access more than one central office line, answer or access a central office line from more than one telephone, and place a line on hold in order to answer or initiate calls on other lines. The system usually includes an intercom capability that allows users on different telephones in the system to communicate with one another. *See* INTERCOMMUNICATING SYSTEMS (INTERCOMS).

The main functional parts of a key telephone system are (1) the central service unit, also called the key service unit (KSU), which serves as an interface and switching center between incoming central office lines and system telephones; (2) station cabling to connect each telephone to the key service unit; (3) multibutton telephone sets, or key telephones, on which separate buttons provide access to the central office lines, intercom, and special features; and (4) a power supply, which converts available electrical power to the specifications required by the key telephone system.

The number of station lines and central office lines served by a key telephone system is limited by the size of the key service unit. On a typical key telephone system, each incoming central office line appears on a separate button on every key telephone set that is allowed to make or receive calls on that line. Usually, a visual indicator, such as a small incandescent lamp or light-emitting diode, allows the user to tell if the line is in use. The same lamp flashes at a certain rate when an incoming call is ringing on the line, and flashes at a different rate when a call is on hold. The number of lines that appear on every key telephone set varies with the capacity of buttons on each set. Such capacities vary from a low of 5 buttons to over 30 buttons, depending on the make and model of the key telephone system.

Technology. Key telephone systems are either electromechanical or electronic in design. The electromechanical type is also called a 1A2 system, or industry-standard key telephone system. The key service unit of an electromechanical system contains relays and electrical components, which amplify the voice signal and switch the signal between the central office lines and station lines. Electromechanical systems are highly reliable and relatively low in cost. However, they require station cabling that contains 25 copper wire pairs, which makes installation complex and changes difficult.

Electronic key telephone systems use microprocessors and integrated circuits to accomplish the switching, voice amplification, and feature activation. The key service unit, as well as the telephones, is proprietary and tends to be more costly than the industry-standard system. However, cabling usually consists of only two to four copper wire pairs, installation is simplified, and the features available are much more extensive and varied than with the 1A2 system. Changes to the system are accomplished by simply reprogramming the microprocessors that control the system. *See* INTEGRATED CIRCUITS; MICROPROCESSOR.

Features. As an option, the key telephone system may allow connection of a music source for playback to callers on hold, and access to centralized facilities such as dictation and paging equipment. Other common features are the ability to transfer a call from one telephone in the system to another, and the ability to unite three or more parties into a single conversation called a conference.

In addition, electronic key telephone systems may offer many specialized features that increase user productivity or meet important business needs. Widely used features include (1) last-number redial, which automatically redials the last number entered by the user; (2) speed dialing, which allows users to call a number either by dialing one or two digits or by depressing a programmed button on the key telephone set; (3) station message waiting, which allows one user to request a return call from another user; (4) hands-free operation, using telephones with built-in speakerphones; (5) station message detail recording, which prints a record of every outgoing telephone call made with the system; and (6) the ability to use inexpensive single-line telephones, in addition to the proprietary multibutton key telephones designed for the system.

Most businesses prefer electronic key telephone systems because of their ease of installation and versatility. Nevertheless, the traditional 1A2 system remains a reliable alternative. *See* SWITCHING SYSTEMS (COMMUNICATIONS); TELEPHONE SERVICE.

Sue R. Rubinstein; Vincent F. Rafferty

Bibliography. The Aries Group · MPSG, *Master Glossary of Terminology*, 1989; Datapro Research, *Telecommunications*, 1990.

Kidney

An organ involved with the elimination of water and waste products from the body. In vertebrates the kidneys are paired organs located close to the spine dorsal to the body cavity and are covered ventrally by the coelomic epithelium. They consist of a number of smaller functional units which are called urinary tubules or nephrons. The nephrons open to larger ducts, the collecting ducts, which open into a ureter. The two ureters run backward to open into the cloaca or into a urinary bladder. The shape and the location of the kidneys varies in different animals. In fish, they are extremely elongated and may reach forward almost to the pericardium. In reptiles, the kidneys are smaller and are located in the pelvic region. In mammals, they are bean-shaped and found between the thorax and the pelvis. The number, structure, and function of the nephrons vary with evolution and, in certain significant ways, with the adaptation of the animals to their various habitats.

Nephron. In its most primitive form, found only in invertebrates, the nephron has a funnel opening into the coelomic cavity followed by a urinary tubule leading to an excretory pore. In amphibians, some of the tubules have this funnel, but most of the tubules have a Bowman capsule (**Fig. 1**). In all higher vertebrates, the nephron has the Bowman capsule, which surrounds a tuft of capillary loops, called the glomerulus, constituting the closed end of the nephron. The inner epithelial wall of the Bowman capsule is in intimate contact with the endothelial wall of the capillaries. Electronmicroscopy reveals that the wall of the capillaries, together with

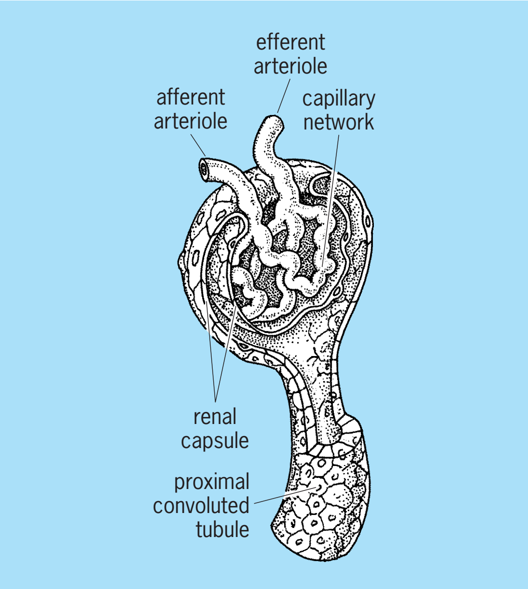


Fig. 1. Nephron, from frog kidney dissected to show glomerulus within Bowman capsule.

the inner wall of the Bowman capsule, forms a membrane ideally suited for filtration of the blood.

The nephron is differentiated into several parts which vary in different groups of vertebrates. Following the glomerulus comes a short neck segment which may be ciliated (this segment is absent in mammals). This is followed by the proximal tubule. Following the proximal tubule in lower vertebrates, there is a short intermediary segment; in mammals, there is the so-called thin segment comprising part of the loop of Henle. This is followed by the distal tubule, which is followed by the collecting duct. In lower vertebrates, all of the nephrons are convoluted and arranged perpendicular to the collecting ducts. In mammals, all of the nephrons have loops of Henle (**Fig. 2**). The glomerulus and convoluted proximal tubule are found in the cortex of the kidney. The loops of Henle run through the outer and inner medulla; some are short, others are long. The thin limb runs into a straight part of the distal tubule, which continues into the distal convoluted tubule found in the cortex. The capillaries are arranged in bundles which form true retia mirabilia, ideally suited for countercurrent exchange. The collecting ducts run from the cortex through the medulla to the tip of the papilla, where they open into the renal pelvis. Through their entire course the collecting ducts are parallel to the loops of Henle and to the capillary loops.

The blood pressure in the capillaries of the glomerulus causes filtering of blood by forcing fluid, small molecules, and ions through the membrane into the lumen of Bowman's capsule. This filtrate contains some of the proteins and all of the smaller molecules in the blood. As the filtrate passes down through the tubule, the walls of the tubule extract those substances not destined for excretion and return them to the blood in adjacent capillaries. Many substances which are toxic to the organism are moved in the opposite direction from the blood into the tubules. The urine thus produced by each nephron is conveyed by the collecting duct and ureter to the cloaca or bladder from which it can be eliminated.

Blood supply. In all classes of vertebrates the renal arteries deliver blood to the glomeruli and through a second capillary net to the tubules. The major blood supply to the kidney tubules comes, however, from the renal portal vein, which is found in all vertebrates except mammals and cyclostomes (Fig. 1). Venous blood from the tail and hindlimbs of the animal drains into the renal portal vein. It then splits up into a second capillary net which surrounds the renal tubules. Waste products from the venous blood can thus be secreted directly into the urinary tubules.

Evolutionary and adaptive aspects. It is assumed that the ancestors of vertebrates were marine forms, with body fluids at osmotic equilibrium with seawater. Migration into fresh water posed the threat of

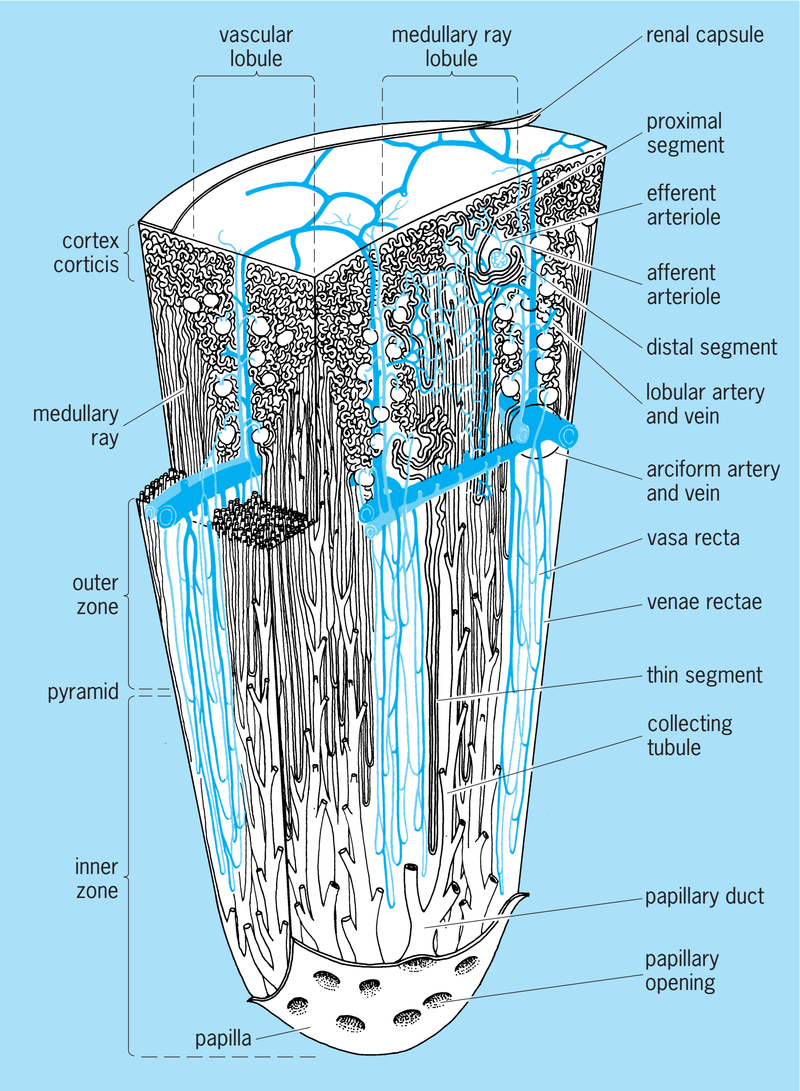


Fig. 2. Diagram of a small section of a human kidney, showing the loops of Henle with their thick and thin segments, and the vasa recta and the collecting ducts as they are arranged parallel to one another in the outer and inner medulla. (*From Homer Smith, Principles of Renal Physiology, Oxford University Press, 1956*)

rapid dilution of body fluids by osmosis. The kidneys in fresh-water forms arose as the major organ for the removal of excess water. A large glomerulus is present in amphibians and fishes found in fresh water today. Their nephrons have a long distal tubule which reabsorbs ions from the filtrate and returns them to the blood. Through this process the urine becomes dilute, and very little salt is lost hereby. However, when fishes reinvaded the ocean, or evolved into land forms, their kidneys had to be modified to their new habitats.

The salt-water teleost fishes maintain an osmotic blood concentration lower than that of the seawater. They are thus in danger of becoming dehydrated, and water conservation becomes a prime objective. Some marine teleosts have lost their glomeruli and the distal tubules entirely, and produce urine by tubular secretion alone. In other marine fishes the glomeruli have been reduced in size.

To maintain osmotic pressure of the blood lower than that of the seawater, the gills secrete salt into the seawater.

In land animals such as reptiles, birds, and mammals, water conservation is also a major objective. The primary function of the kidneys is to excrete salt and nitrogenous waste products. Reptiles and birds save water by excreting their nitrogenous wastes in the form of a nearly dry paste of uric acid crystal. The kidney of reptiles can only make urine that is isoosmotic or hypoosmotic to the blood, but not hyperosmotic. Reptiles faced with a large salt intake, such as marine turtles and lizards, excrete salt by special salt glands in the head. Birds can produce a urine hyperosmotic to the blood because of the countercurrent system formed by the few loops of Henle. They excrete uric acid. Salt-water birds excrete excess salt through glands similar to those of the reptiles.

Mammals are peculiar in that, regardless of habitat, they always excrete almost all of their nitrogenous wastes in the form of urea. It may be for this reason that the mammal has developed its unique kidney, which can produce urine with several times the osmotic concentration of the blood. The ability of the mammalian kidney to concentrate urine is inherent in the spatial arrangement of renal tubules, blood vessels, and collecting ducts which permits the tubules to function together as a countercurrent multiplier system (Fig. 2). *See* EXCRETION; OSMOREGULATORY MECHANISMS.

Embryology. Although nephrons are assembled into kidneys in a variety of ways, the tubules always arise in the intermediate mesoderm along the upper border of the body cavity (**Fig. 3**). The first tubules develop at the anterior end of this mesoderm, followed by others in sequence behind. Variations in this plan permit the following kidney types to be distinguished: holonephros, pronephros, mesonephros, opisthonephros, and metanephros.

Holonephros. There is one nephron beside each somite along the entire length of the coelom. Supposedly, this is the ancestral condition and is thus referred to as an archinephros. It is seen today only in the larvae of a few primitive fishes, the myxinoid cyclostomes.

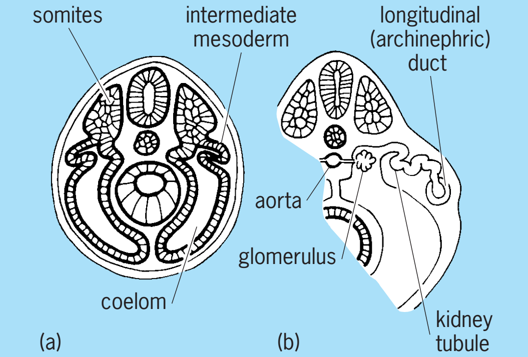


Fig. 3. Diagrammatic cross section of the embryos to show origin of nephron. (*a*) Early stage with mesoderm differentiation. (*b*) Later stage.

Pronephros. Nephrons are formed only in the anterior part of the intermediate mesoderm. The pronephros is rarely found as an adult excretory organ, but is important because in all embryos it initiates the formation of the archinephric duct used by subsequent tubules for drainage. The pronephros is of interest also because it shows features presumed to be primitive, suggesting a series of stages in evolution of the nephron.

Mesonephros. The nephrons form behind the pronephros along the middle zone of the coelom. The tubules usually become involved in sperm transport. Drainage is through the archinephric duct. The mesonephros never forms an entire adult kidney. It is functional in the chick and other amniote embryos.

Opisthonephros. Nephrons develop from the middle and posterior portions of intermediate mesoderm. The middle, or mesonephric, tubules sometimes form only a minor part. The archinephric duct may be supplemented or even replaced by ureterlike accessory drainage ducts. This is the functional adult kidney in amphibians and fishes.

Metanephros. Nephrons form from only the extreme posterior part of the intermediate mesoderm, and are drained by a distinctive new canal, the ureter, which grows forward from the base of the archinephric duct before the tubules have begun to develop. This is the kidney of adult reptiles, birds, and mammals. Their embryos show a distinct sequence of pro-, meso-, and metanephric kidneys. *See* URINARY SYSTEM.

Bodil Schmidt-Neilsen

Kidney disorders

Disorders of the kidney are classified by both their etiology and their anatomic location along the nephron. Disorders other than those arising from tumors become evident when the kidney is unable to regulate the volume and composition of the extracellular fluids, resulting in edema and hypertension. The clinical signs cannot identify the specific segment of the nephron that is affected, since overall renal function depends on each segment functioning in sequence. The failure of one part, therefore, affects the functioning of the whole unit and makes a precise diagnosis difficult without a kidney biopsy. Knowing the location and type of abnormality is important because it may determine the prognosis or therapy.

Kidney disorders may be divided into the following categories: congenital, inflammatory, hereditary, infection-related, or associated with metabolic diseases (such as diabetes mellitus), toxins or drugs, circulatory collapse, and cancer.

Congenital abnormalities. The absence of kidneys at birth, which is quite rare, is incompatible with life. A slightly more common condition, however, is the presence of only one kidney or the presence of one kidney in an abnormal location.

The most common congenital disorders are abnormalities in the peristaltic passage of urine. Normally, urine flows from the kidneys into the ureters, where it is moved by a rhythmic peristaltic motion to the bladder. It is stored there until voluntary voiding is begun, whereupon it passes through the urethra and out of the body. Any abnormality in such a complex and integrated neuromuscular system responsible for the passage of urine or any mass of tissue that blocks urine flow causes retention of the urine. All such abnormalities are referred to as obstructive lesions, even though some are not actual physical masses. The most common obstructive lesions are tissue masses that act as valves, inhibiting passage of urine in the normal direction but allowing retrograde flow. In many cases, diseases that inhibit the normal peristaltic flow of urine down the ureters because of a neuromuscular abnormality gradually run their course and do not require surgery. On the other hand, "valves" must be promptly removed surgically. *See* CONGENITAL ANOMALIES.

Hereditary disorders. The most common hereditary kidney disorder is adult polycystic kidney disease. It is an autosomal dominant disease that is found in all racial and ethnic groups and is characterized by the formation of cysts along the length of the nephron. As the cysts enlarge, the kidneys likewise enlarge—resulting in kidney failure in midadulthood. Approximately 10% of patients who undergo dialysis have polycystic kidney disease.

Alport's syndrome, a very rare genetic disease of the glomeruli, was first described in Finland and now is known to occur throughout the world. The disease results in glomerular scarring and eventual renal failure within the second or third decade of life that cannot be reversed. *See* HUMAN GENETICS.

Infection-related diseases. The kidneys are subject to infection from both local and remote sources and from a variety of causative organisms.

Pyelonephritis. The most common cause of pyelonephritis, a bacterial infection of the kidney, is obstruction of the bladder or ureters by a tumor or kidney stones that interrupt urine flow. The static urine allows bacteria to ascend to the kidney from the perineum, which accounts for the fact that the organisms most frequently involved are coliform bacteria, which are normal residents of that area. Treatment consists of removal of the obstruction and antibiotic therapy. Bacteria may also reach the kidney through the bloodstream from an infection elsewhere, but that is much less common.

Glomerulonephritis. The kidney may also be subject to infections originating at distant sites. Examples include bacterial products, such as endotoxins, that are released into the bloodstream and antibodies that are formed in response to bacterial invasion. In the latter case, the antibodies form antigen-antibody complexes in the circulation or the kidney. The complexes that are localized in the kidney cause an inflammatory reaction in the glomeruli, a process called acute glomerulonephritis. It is most frequently associated with streptococcal infections of the skin or the upper respiratory tract and is common in children, who frequently develop streptococcal infections. It is usually a short-lived ailment, and nearly all patients recover completely. *See* STREPTOCOCCUS.

Metabolic disorders. The most common metabolic disorder to affect the kidney is diabetes mellitus. Of those persons with the insulin-dependent form of diabetes, 30–40% develop renal disease. Kidney dysfunction is one of the most life-threatening complications of diabetes: between the ages of 25 and 40, the mortality rate of diabetics with kidney diseases is nearly 20 times that of the general population. Diabetes causes progressive glomerular scarring, which leads to kidney failure. No treatment has yet been devised to prevent either the onset or progression of diabetic kidney disease; however, careful control of blood sugar and systemic blood pressure are probably important. *See* DIABETES.

Systemic diseases. The two most common systemic diseases to affect the kidneys are systemic lupus erythematosus and vasculitis. Systemic lupus erythematosus is essentially a disease of young women and is characterized by the deposition of immunoglobulins and other immune substances in the glomeruli. The resulting acute inflammatory reaction may lead to progressive scarring. Systemic lupus erythematosus often responds to treatment with immunosuppressive agents. *See* CONNECTIVE TISSUE DISEASE.

Systemic vasculitis can affect either the medium-sized arteries or the arterioles. Involvement of the former leads to infarcts in various organs but seldom causes kidney failure. When the arterioles are affected, the lesions are diffuse, because the kidney is highly vascular and the arterioles, the smallest of the arteries, are abundant. Kidney failure often occurs. The cause of the disease is unknown, but those who have vasculitis often show systemic signs, including fever. The vascular lesions are characterized by an acute inflammatory infiltrate that involves the blood vessel wall. Treatment with anti-inflammatory medications can help clear the lesions.

Toxins and drugs. The kidney is a major site for the excretion of waste products as well as heavy metals, the acids and alkalies of metabolism, and drugs. Because the urine becomes concentrated (100-fold) as it travels the length of the tubule, the concentration of otherwise-innocuous, or even therapeutic, substances may reach levels that are toxic to the kidneys. The accumulation of toxins within the tubule often leads to necrosis, or cell death, of the tubular epithelium or an inflammatory reaction in the spaces between the tubules. The result in either case is acute kidney failure. Usually, removing the offending agent and relying on dialysis until renal function returns reverses the condition. Chronic exposure to some phenacetin-containing compounds, however, may lead to progressive kidney fibrosis, which appears slowly and is irreversible. *See* TOXICOLOGY.

Circulatory collapse. If systemic blood pressure drops to levels at which the kidneys are no longer perfused, the tubules become deprived of oxygen and suffer severe injury. The most common causes of underperfusion are blood loss and cardiac failure. If bloodflow to the kidneys is restored, epithelial cell damage is rapidly reversed and renal function returns to normal. *See* VASCULAR DISORDERS.

Cancer. Cancer of the kidney occurs mainly in two age groups—the very young and those over 50 years of age. Tumors found in the young, the most common type being Wilms' tumor, are often bilateral, composed of fetal tissues, and may grow to a large size before detection. Some cancerous tumors of the kidney in children are responsive to radiation or chemotherapy.

Tumors in the adult kidney, adenocarcinomas or hypernephromas, are formed of tubular cells and usually metastasize to the lungs and bones. Blood in the urine is a common early sign, and if all traces of the cancerous growth are removed surgically, a cure is possible. *See* CANCER (MEDICINE); KIDNEY; URINARY SYSTEM.

Gary Striker

Bibliography. S. Massry and R. Glassock (eds.), *Textbook of Nephrology*, 4th ed., 2000; P. Sweny, *The Kidney and Its Disorders*, 1989; P. Sweny and Z. Varghese, *Renal Disease*, 1988; C. Tisher and B. Brenner (eds.), *Renal Pathology*, 2d ed., 1994.

Killing equations

A set of special, linear, first-order partial differential equations involving the coordinate variables x of an N-dimensional riemannian space, the metric tensor g_{ab} of this space, and a vector whose components V_a, V^a ($V^a = g^{ab}V_b$) are functions of x. Specifically, it is given by either Eq. (1) or Eq. (2). The subscript semicolon indicates partial differentiation with respect to a coordinate variable x, while the comma specifies covariant differentiation. The equivalence of Eqs. (1) and (2) may be established as follows. From the definitions of covariant differentiation and the structure of the Christoffel symbols { }, Eq. (3) follows, and one obtains Eq. (4) by virtue of the symmetry of g_{ab}.

$$V^c g_{ab;c} + g_{ac}V^c{}_{;b} + g_{bc}V^c{}_{;a} = 0; \tag{1}$$

$$c \equiv \frac{\partial}{\partial x^c}$$

$$V_{a,b} + V_{b,a} = 0 \tag{2}$$

$$\begin{aligned} V_{a,b} &= V_{a;b} - V_{d\{{}^{d}_{ab}\}} \\ &= (V^c g_{ac})_{;b} - V_d g^{cd^{1/2}}[-g_{ab;c} + g_{cb;a} + g_{ac;b}] \end{aligned} \tag{3}$$

$$\begin{aligned} V_{a,b} + V_{b,a} &= (V^c g_{ac})_{;b} + (V^c g_{bc})_{;a} \\ &\quad -{}^1/_2 V^c[-g_{ab;c} - g_{ba;c} + g_{cba;a} \\ &\quad + g_{ca;b} + g_{ac;b} + g_{bc;a}] \\ &= g_{ac}V^c{}_{;b} + g_{bc}V^c{}_{;a} + V^c g_{ab;c} \end{aligned} \tag{4}$$

The tensor components $V_a(x)$ are supposedly of the type (0,1, 0). Specifically, transformation equation (5) for V_a is induced by the change of coordinates $x^a = x^a(\overline{x})$, $\overline{x}^r = \overline{x}^r(x)$. The quantities $V_{a,b}$ obtained by covariant differentiation are tensor components of the type (0,2, 0). Specifically, they obey Eq. (6), and therefore Eq. (7) is valid. Hence if for a given tensor V_a [type (0,1, 0)] and a certain coordinate system (x) of the infinite collection of tensor analysis the Killing equations are valid, then these equations are valid for each coordinate system of the infinite collection. Briefly, the Killing equations are invariant for class C' coordinate transformations.

$$\overline{V}_r = V_a \frac{\partial x^a}{\partial \overline{x}^r} \tag{5}$$

$$\overline{V}_{r,s} = V_{a,b} \frac{\partial x^a}{\partial \overline{x}^r} \frac{\partial x^b}{\partial x^s} \tag{6}$$

$$\begin{aligned} \overline{V}_{r,s} + \overline{V}_{s,r} &= V_{a,b} \frac{\partial x^a}{\partial \overline{x}^r} \frac{\partial x^b}{\partial \overline{x}^s} + V_{b,a} \frac{\partial x^b}{\partial \overline{x}^s} \frac{\partial x^a}{\partial \overline{x}^r} \\ &= (V_{a,b} + V_{b,a}) \frac{\partial x^a}{\partial \overline{x}^r} \frac{\partial x^b}{\partial \overline{x}^s} \end{aligned} \tag{7}$$

The Killing equations are fundamental with regard to the question of the existence of certain mappings (or, synonymously, transformations) of a space onto itself. These particular transformations are frequently called infinitesimal groups of motions and are related to the ordinary groups of motions, that is, the motions which have the characteristic group properties.

Groups of motions. The one-parameter motions determined by mapping equations of the type in Eq. (8) will be said to constitute a one-parameter group of motions if the following group properties hold: (1) there exists an admissible value of t, t_0, such that for each admissible x, $x^{*a}(t_0) = f^a(x, t_0) = x^a$ (the identity transformation); (2) for each admissible set of x-values x^{*a} and each admissible value for t_1, the equations $x^{*a} = f^a(x, t_1)$ have a unique solution expressible in the form $x^a = f^a(x^*, t_2)$ [existence of an inverse]; (3) for each admissible set of x-values and each admissible pair of t-values t_1, t_2, there exists an admissible parameter value t_3, such that $x^{**a} = f^a(x^*, t_2) = f^a[f(x, t_1)t_2] = f^a(x, t_3)$ [the closure or principal group property]. The term "groups of motions" has been used at times under lighter conditions than

$$x^{*a} = f^a(x, t) \qquad (x \equiv \text{set } x^1, x^2, \ldots, x^N) \tag{8}$$

these; for example, condition 1 or 2, or both, has been omitted. *See* GROUP THEORY.

A very simple example of a group of motions for an ordinary euclidean plane E_2 and a rectangular cartesian coordinate system is given by Eq. (9), with t

$$x^{*a} = f^a(x, t) = x^a + t \qquad (9)$$

restricted to real numbers. Group property 1 is obviously met by $t = 0$. Property 2 is satisfied since the inverse of $x^{*a} = x^a + t_1$ for each admissible value of t_1 is $x^a = x^{*a} + (-t_1) = f^a(x^*, -t_1)$. Property 3 and Eq. (9) require that for each admissible choice of t_1 and t_2 there exists an admissible t-value t_3 such that $x^{**a} = f^a(x, t_3) = x^a + t_3$ with $x^{**a} = f^a(x^*, t_2) = x^{*a} + t_2 = (x^a + t_1) + t_2$. Obviously the requirement of property 3 is fulfilled by $t_3 = t_1 + t_2$. Thus the group requirements are satisfied. The motion criterion is also met since $g_{ab} = \delta_{ab}$ ($= 1$ for $a = b$, $= 0$ for $a \neq b$) throughout the plane. Thus a change in the evaluation point does not cause a change in the value of g_{ab}, and the metric magnitudes (such as arc lengths and angles between intersecting curves) are preserved. A slight generalization of the motion of Eq. (9) is given by Eq. (10), with k^1 and k^2 given constant real numbers, not both zero.

$$f^a(x, t) = x^a + k^a t \qquad (10)$$

As another simple example of a motion, consider a plane E^2 which has a point O deleted and contains a rectangular cartesian coordinate system (x) with center at O; and impose the point transformation in Eqs. (11). The identity transformation is given by

$$x^{*1} = f^1(x, \theta) = \cos\theta\, x^1 - \sin\theta\, x^2 \qquad (11a)$$

$$x^{*2} = \sin\theta\, x^1 + \cos\theta\, x^2 \qquad (11b)$$

$$0 \leq \theta < 2\pi$$

$\theta = 0$; and for $0 < \theta_1 < 2\pi$, the inverse is $x^a = f^a(x^*, \theta_2)$, $\theta_2 = 2\pi - \theta_1$. The group requirement is satisfied provided there is a functional relation $\theta_3 = \theta_3(\theta_1, \theta_2)$ for $0 \leq \theta_1 < 2\pi$, $0 \leq \theta_2 < 2\pi$, the relation having the following properties (1): $0 \leq \theta_3 < 2\pi$; (2) from $x^{**a} = f^a(x^*, \theta_2)$, $x^{*a} = f^a(x, \theta_1)$, it follows that $x^{**a} = f^a(x, \theta_3)$. From the rotational interpretation of mapping equations (11), it is evident that $\theta_3(\theta_1, \theta_2)$ may be defined by $\theta_3(\theta_1, \theta_2) = \theta_1 + \theta_2$ for $0 \leq \theta_1 + \theta_2 < 2\pi$; and $\theta_3(\theta_1, \theta_2) = \theta_1 + \theta_2 - 2\pi$ for $\theta_1 + \theta_2 \geq 2\pi$. Also, $g^*_{ab}(dx^{*a}du)\,(dx^{*b}dv) = g_{ab}(dx^a du) \cdot (dx^b dv)$, and it is evident that the metric entities are preserved in value by the mapping.

Formulation of motions. The definitive requirement for a mapping to be a motion, namely, the preservation of the magnitudes of the metric entities, is of course very stringent. Given a riemannian space R, with a coordinate system (x), one method of attack for searching for a motion is to seek a suitable coordinate transformation which preserves the structure of the metric tensor. If such a coordinate transformation $\bar{x}^r = \bar{x}^r(x)$ exists, then the tensor transformation in Eq. (12) can be altered in structure by replacing

$$g_{ab}(x) = \bar{g}_{rs}[\bar{x}(x)]\frac{\partial \bar{x}^r}{\partial x^a}\frac{\partial \bar{x}^s}{\partial x^b} \qquad (12)$$

$\bar{g}_{rs}[\bar{x}(x)]$ with $g_{rs}[\bar{x}(x)]$ since the functional structure of $\bar{g}_{rs}(\bar{x})$ is the same as that of $g_{rs}(x)$ and therefore $\bar{g}_{rs}(\bar{x}) = g_{rs}(\bar{x})$. If this change is made, Eq. (13) results.

$$g_{ab}(x) = g_{rs}[\bar{x}(x)]\frac{\partial \bar{x}^r}{\partial x^a}\frac{\partial \bar{x}^s}{\partial x^b} \qquad (13)$$

Conversely, if the coordinate transformation is not known, then this last equation may be regarded tentatively as a set of $N(N + 1)/2$ first-order partial differential equations in N unknown functions $\bar{x}^r(x)$. While motions can be formulated in very special cases, the desirability of replacing the motion requirement with another less demanding one is obvious.

Modified motion requirement. If C_1 and C_2 are two class C' arcs given by $x^a = x_1{}^a(u)$, $x^a = x_2{}^a(v)$ which intersect at a point P for $u = u_0$, $v = v_0$ and if, in addition, the motion requirement is met, then $g_{ab}(x^*) \cdot dx_1{}^{*a}/du\,dx_2{}^{*b}/dv|u_0, v_0$ is a constant function of t and therefore its derivative vanishes identically in t. A natural reduction in the severity of this requirement is obtained by imposing the alternate condition that the derivative vanish at t_0, the identity value for t, which is now assumed to be $t_0 = 0$. It follows that $f^s(x, 0) = x^a$ and hence, by the McLaurin expansion for $f^a(x,t)$ that $x^{*a} = x^a + t \cdot [(\partial f^a/\partial t)|t = 0] + \cdots$. These facts suggest consideration of the simpler point transformation, Eq. (14). For each fixed

$$x^{*a} = x^a + tV^a(x) \qquad (14)$$

point $P(x)$, Eq. (14) represents a parametrized arc through P, and the tangent vector at P has the components $\partial x^{*a}/\partial t|P$ which are equal to $V^a(x)$ at P. In the literature Eq. (14) is frequently written in the modified form of Eq. (15), and referred to as an infinitesimal transformation.

$$x^{*a} = x^a + \delta t V^a(x) \qquad (15)$$

With regard to the motion requirement: (1) for a parametrized arc C given by $x^a = x^a(u)$ and of class C', $dx^{*a}/du = dx^a/du + V^a{}_{;b}(dx^b/du)\, t$ and therefore $\partial/\partial t(dx^{*a}/du) = V^a{}_{;b}dx^b/du$; and (2) $g_{ab}|MP = g_{ab}(x^*) = g_{ab}(x^c + V^c t)$ and hence $\partial g_{ab}/\partial t = \partial g_{ab}/\partial x^{*c}V^c$, and $\partial g_{ab}/\partial t_{t=0} = g_{abc}(x)V^c$. Hence for the case of the two arcs C_1 and C_2, previously introduced, Eq. (16) follows. Thus the Killing equations

$$\left.\frac{d\left[g_{ab}(x^*)\dfrac{dx_1{}^{*a}}{du}\dfrac{dx_2{}^{*b}}{dv}\right]}{dt}\right|_{t=0}$$

$$= g_{ab;c}(x)V^c\frac{dx_1{}^a}{du}\frac{dx_2{}^b}{dv} + g_{ab}(x)V_{;c}{}^a\frac{dx_1{}^c}{du}\frac{dx_2{}^b}{dv}$$

$$+ g_{ab}(x)\frac{dx_1{}^a}{du}V_{;c}{}^b\frac{dx_2{}^c}{dv}$$

$$= (g_{ab;c}V^c + g_{cb}V_{;a}{}^c + g_{ac}V_{;b}{}^c)\frac{dx_1{}^a}{du}\frac{dx_2{}^b}{dv} \qquad (16)$$

constitute a necessary and sufficient condition for a point transformation, Eq. (14), to be an infinitesimal motion. *See* DIFFERENTIAL EQUATION; RIEMANNIAN GEOMETRY; TENSOR ANALYSIS. Homer V. Craig

Bibliography. I. Chavel, *Riemannian Geometry: A Modern Introduction*, 1995; L. P. Eisenhart, *Riemannian Geometry*, rev. ed., 1950, reprint, 1997; W. P. A. Klingenberg, *Riemannian Geometry*, 2d ed., 1995.

Kiln

A device or enclosure to provide thermal processing of an article or substance in a controlled temperature environment or atmosphere, often by direct firing, but occasionally by convection or radiation heat transfer. Kilns are used in many different industries, and the type of device called a kiln varies with the industry.

"Kiln" usually refers to an oven or furnace which operates at sufficiently high temperature to require that its walls be constructed of refractory materials. The distinction between a kiln and a furnace is often based more on the industry than on the design of the device. For instance, an electrically heated refractory tunnel oven equipped with a stainless mesh conveyor belt to carry the work through is referred to as a tunnel kiln if it is used for sintering small ceramic electronic parts such as ferrite transformer cores. The same device used to sinter small metal parts from powdered aluminum alloys is called a sintering furnace.

Generally the word "kiln" is used when referring to high-temperature treatment of nonmetallic materials such as in the ceramic, the cement, and the lime industries. When melting is involved as in steel manufacture, the term "furnace" is used, as in blast furnace and basic oxygen furnace. In glass manufacture, the melting furnace is often called a glass tank when the process is continuous. *See* FURNACE CONSTRUCTION.

Rotary kiln. This is the largest type of kiln, being used for heating of loose bulk materials such as cement and lime. Rotary kilns are used exclusively for the production of cement and for 85% of the commercial quicklime produced in the United States. Over the years, rotary cement kilns have increased in size to gain greater capacity and productivity from a single unit: kilns having a shell as large as 21 ft (6.3 m) diameter by 700 ft (210 m) long have been constructed. This is probably approaching the practical size limit. *See* CEMENT; LIME (INDUSTRY).

A rotary kiln (**Fig. 1**) is a long, refractory-lined cylinder supported on steel hoops, or "tires," riding on rollers, or "trunnions." The kiln is inclined at a slight slope, often $^{1}/_{2}$–$^{5}/_{8}$ in. fall per foot (42–52 mm per meter), and rotated at a slow speed (about 0.8–1.0 rpm for cement and lime kilns). For calcining ores, there is a tendency to use flatter slopes and higher rotational speeds because of the improved heat transfer obtained. A burner (usually using pulverized coal, natural gas, fuel, oil, or a combination of fuels) is placed at the lower end to supply the necessary heat. Raw feed enters the upper end and slowly tumbles through the kiln, being heated by the countercurrent flow of hot combustion gases. Physical and chemical changes take place in the charge, such as drying and dehydration, decomposition of carbonates, and fusing of materials into a clinker.

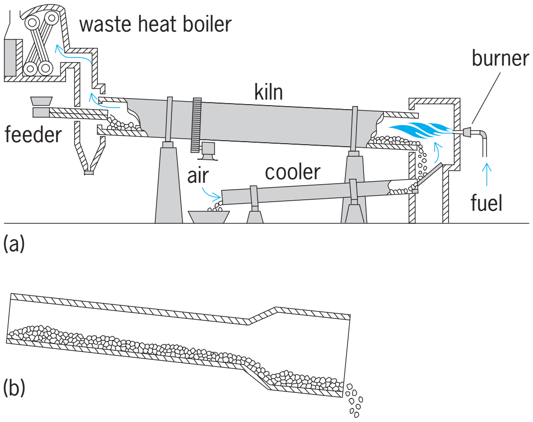


Fig. 1. Rotary kilns: (*a*) single-diameter; (*b*) two-diameter. (*After R. H. Perry and C. H. Chilton, Chemical Engineers' Handbook, 5th ed., McGraw-Hill, 1973*)

A rotary kiln tends to have poor thermal efficiency. To conserve fuel, auxiliary devices are usually added. The hot product drops into a cooler which preheats the combustion air. Chain curtains at the feed end and feed-preheating devices are frequently used to extract heat from the flue gas and preheat the kiln feed.

Rotary kilns are also used for calcining or agglomerating phosphate rock; alumina and bauxite; magnesia; clays; carbon and petroleum coke; lightweight aggregates; iron, chromium, lithium, and uranium ores; burning of elemental sulfur; and incineration or pyrolysis of combustible wastes.

Shaft kiln. A shaft kiln (**Fig. 2**) consists of a vertical, stationary, refractory-lined steel tube, frequently tapered to grow slightly larger toward the bottom to permit free downward movement of the charge. These kilns have been used for centuries for calcining limestone and cement and are still used worldwide (except in the United States) for cement production; 15% of the commercial quicklime made in the United States is made in shaft kilns. Shaft kilns are now also used for calcining taconite pellets and as an oil shale retort. The raw charge is added at the top and is discharged through grates at the bottom. Originally, coal or coke was mixed with the charge to supply fuel and was burned out by hot air introduced through the walls above the grates.

Most shaft kilns are now equipped with external burners to supply the heat in order to eliminate product contamination caused by in-place firing of solid fuels. Shaft kilns have much better heat economy than rotary kilns since the gases flowing up through the kiln have more intimate contact with the solid charge. However, shaft kilns generally have lower production rates than rotary kilns, and accurate control of the charge time-temperature relationship is less precise because of gas channeling.

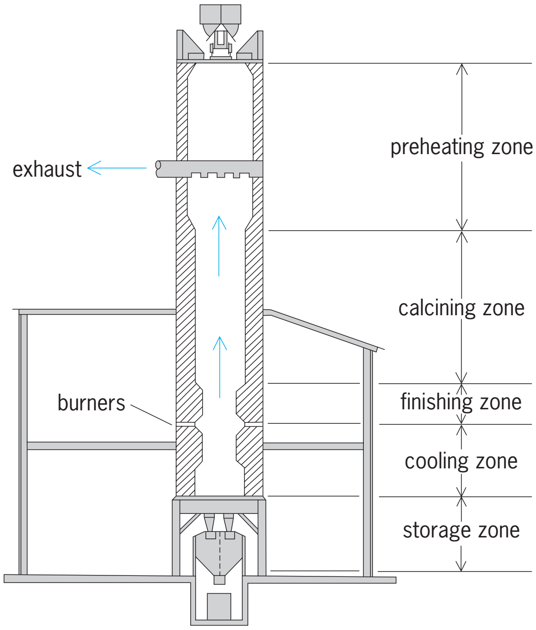


Fig. 2. Vertical shaft kiln. (*After C. J. Lewis and B. B. Crocker, The lime industry's problem of airborne dust, J. APCA, 19:31–39, 1969*)

Newer lime kilns. Two developments for lime calcining are the fluo-solids lime kiln, in which the limestone is calcined in a bed fluidized with hot combustion products, and the Calcimatic kiln, which is a horizontal rotating, circular refractory hearth kiln. The charge is deposited continuously on a pie-shaped element of the hearth as the hearth rotates about its center. The charge advances below radiant burners on the underside of a refractory dome roof. Rotational speed is set so that the charge is calcined when the hearth has rotated about 350°, at which point the charge is scraped from the hearth with a conveyor and transferred to a shaft cooler which preheats combustion air. One advantage of the Calcimatic is the low dust emission in the flue gas.

Kilns for formed ceramic products. Several types of kilns, either batch or continuous, are used for firing formed ceramic products. An early batch kiln is the scove kiln, in which unfired bricks are stacked to provide fireboxes and flues. Thus the basic material to be fired forms the kiln structure, which is covered on the outside with fired brick and a clay wash. Such kilns holding 30,000–200,000 bricks were fired with wood to produce all early American brick. Later, coal or oil fuel was used. *See* CERAMICS.

Periodic kiln. This is the general name given to a batch-type kiln in which the ware passes through a cycle consisting of heat-up, soaking (holding at peak temperature for some time), cooling, and removing or "drawing" the ware at the end of the cycle (**Fig. 3**). Such kilns vary in size from small laboratory or art-pottery kilns to large domed structures 40–50 ft (12–15 m) in size. Periodic kilns have poor thermal efficiency since all the heat used to heat the ware is lost during cooling. Such units are also wasteful of worker-power in loading and unloading.

Another early kiln, the beehive kiln, is a type of periodic kiln. It consists of a permanent, hemispherical chamber surrounded by a number of fuel beds, often burning coal. The hot gases pass up inside the walls and down through the material to be fired to a central flue in the floor.

Chamber kiln. This kiln, widely used in Europe but not generally adopted in the United States, is one of the earliest attempts to improve heat efficiency and approach continuous firing. The kiln is built in the form of a ring or long oval, divided into separate chambers connected with flues. All chambers are loaded with ware. Firing takes several days, the fire being started in one chamber and then moved progressively around the circuit. The combustion gas passes through chambers yet to be fired, where the ware is preheated. After completion of firing in a chamber, combustion air is passed through to cool the ware and preheat itself.

Tunnel kilns. These are long structures in which the ware is moved through the kiln (**Fig. 4**). Combustion air is introduced countercurrent to the unfired ware. Thus the air is preheated by cooling the fired ware. Fuel is burned in the center section, and the hot combustion gases preheat the incoming ware and remove its moisture. Tunnel kilns may be as small as a few inches wide and high by 15–30 ft (4.5–9 m)

Fig. 3. Interior of periodic kiln, with electrical porcelain insulators for firing. (*Swindell-Dressler, Inc.*)

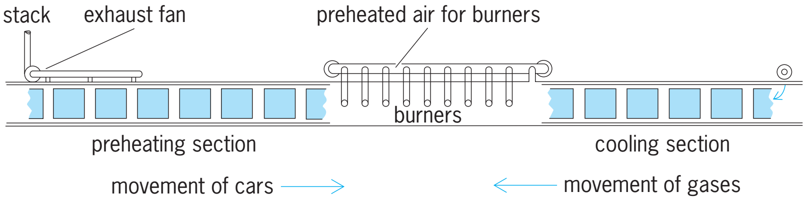


Fig. 4. Section of direct-fired tunnel kiln showing principle of operation. (*After T. J. W. van Thoor, ed., Chemical Technology: An Encyclopedic Treatment, vol. 2, Barnes and Noble, 1971*)

long for small ceramic parts to 6 by 8 ft (1.8 by 2.4 m) in cross section and several hundred feet long for continuous brick processing.

A tunnel kiln is often named after the method of moving the work through the kiln. Brick is loaded on a refractory-topped small car which rolls through the tunnel on a track. The refractory hearth of the car forms the bottom of the kiln. Such kilns are called car tunnel kilns. The car is frequently pushed into the tunnel with a hydraulic ram. Movement of the cars may be either continuous or intermittent at spaced time intervals. Small tunnel kilns may have a traveling mesh belt (mesh belt tunnels), or the ware may be placed in metal or refractory boxes called saggers. The boxes may be pushed through on a stationary hearth (pusher tunnel) or may roll on top of ceramic rollers (roller hearth tunnel).

Muffle kiln. In this type of kiln, which may be either periodic or continuous, the ware is protected from direct contact with the fire by a radiant refractory wall if the kiln is fuel-fired. In many special small muffle tunnels, electric heat is used. The tunnel refractory lining may be wound with a chrome-nickel resistance element, or silicon carbide glowbar units may be suspended from the roof and buried in the hearth. A muffle kiln may be used for firing high-priced dinnerware. Special electrically heated muffle tunnels have been developed for the electronics industry to fire ceramic capacitors, ceramic magnets, and printed ceramic circuits, often with carefully controlled atmospheres such as dry hydrogen.

Lumber and tobacco kilns. Kilns used in these industries are basically low-temperature dryers and do not require refractories for their construction. Lumber-drying kilns are of two basic types: a box or compartment type in which the lumber is loaded in batches for drying, and a progressive kiln in which the lumber travels through on cars much as in a tunnel kiln. The lumber is stacked horizontally with “stickers” (uniform wood spacers) between successive layers. Air of controlled temperature and humidity is passed over and through the lumber, often with circulating fans. Older kilns may have natural circulation produced by rising currents of hot air from steam heating coils. Fan kilns have better thermal efficiency because the air is recirculated more times before venting. Air temperatures vary from 150 to 200°F (66 to 93°C), depending on lumber size and species.

A tobacco-drying kiln is an airtight barn used for flue-curing of tobacco. Leaf strings of tobacco are hung from racks. Wood, coal, oil, or natural gas is burned in a firebox, or “kiln.” The hot flue gases circulate through sheet-iron flues near the floor of the barns. The leaf is subjected to a carefully controlled temperature-time cycle (usually 3–5 days long) to produce the desired characteristics in the leaf.

Heating. Gas and oil have been the preferred fuels for all types of kilns and will probably remain so for moderate-priced ceramics. Pulverized coal has been used in the past for firing rotary kilns and will probably become the preferred fuel of the future because of the scarcity of hydrocarbon fuels. Coal has been burned on an external grate to fire shaft kilns when oil or gas is not available. Coal gas and producer gas have also been generated for this purpose. Electric heat is generally limited to smaller kilns processing valuable wares where precise temperature control is needed or where convenience is important, as in laboratory and pilot plant kilns.

Pollution control. Kilns can be sources of air pollution arising either from the combustion of the fuel or from entrainment of dust particles from the charge. The latter source applies only to kilns where loose bulk materials are processed, such as rotary and shaft kilns. The exhaust gases from these kilns are relatively hot, and cyclonic collectors have often been used to remove dust particles down to 5–10 micrometers in size. This collection does not meet today’s requirements, and such kilns are being equipped with electrostatic precipitators or bag filters. Wet scrubbing can be used efficiently, but this results in a condensing steam plume in cool weather, which may be esthetically objectionable.

Combustion pollution can take the form of incompletely burned fuel (hydrocarbons, carbon monoxide, soot, and smoke) or oxides of nitrogen. The former is prevented by good combustion practices—adequate excess air, proper mixing of fuel and air, and adequate combustion temperatures. Nitrogen oxides are produced only when combustion is carried out at very high temperatures (usually 2000–2800°F or 1093–1538°C) to produce high peak temperatures in the kiln. NO_x is controlled by reducing flame temperature, using less excess air, two-stage combustion, and recirculation of exhaust gases. *See* AIR POLLUTION; DUST AND MIST COLLECTION.

Burton B. Crocker

Bibliography. L. Culpepper, *High Temperature Drying: Enhancing Kiln Operations*, 1990; F. Olsen, *The Kiln Book*, 3d ed., 2001; K. Peray, *Rotary Cement Kiln*, 2d rev. ed., 1986; R. H. Perry and D. Green (eds.), *Perry’s Chemical Engineers’ Handbook*, 7th ed., 1997; W. Trinks and W. H. Mawhinney, *Industrial Furnaces*, vol. 1, 5th ed., 1961, vol. 2, 4th ed., 1967.

Kimberlite

A variety of peridotite, an igneous rock containing at least 35% olivine. The full definition is unusually complex for an igneous rock, with many qualifications. Briefly, kimberlite is richer in carbon dioxide than most peridotites, and has crystals larger in diameter than 0.5 mm of olivine, garnet, clinopyroxene, phlogopite, and orthopyroxene. All of these silicate minerals have high Mg/Fe ratios in kimberlites. In addition, there can be 0.5 mm or larger crystals of magnesian ilmenite and titanium-poor chromite. The matrix (crystals less than 0.5 mm) contains olivine plus one or more of monticellite, phlogopite, perovskite, spinel, apatite, serpentine, and carbonate minerals. Melilite, leucite, kalsilite, and alkali feldspar do not crystallize from kimberlite magma. In part, the complexity of this definition is caused by the complicated processes of generation and emplacement of kimberlites and, in part, a restrictive definition is imposed as a defense against wishful thinking, because kimberlite was long thought to be the only rock that brought diamonds to the Earth's surface. However, in the 1970s other diamond-bearing rocks (lamproites) were found in western Australia and elsewhere, and in the 1980s R. H. Mitchell distinguished orangeite and kimberlite as separate hosts for diamonds in South Africa. Some eclogites have recently been found to contain diamonds also. *See* ECLOGITE; OLIVINE; PERIDOTITE.

Diamonds are the only economically significant mineral extracted from kimberlite. They form deeper than 150 km in the Earth's mantle and are carried upward as "accidental tourists" in kimberlite. Several thousand kimberlite occurrences are known, but less than 10% of these contain any diamonds, and less than 1% can support profitable mines. *See* DIAMOND.

The magmatic liquid that forms kimberlite is generated by the melting of small amounts of the Earth's upper mantle containing water and carbonate. The liquid moves upward, gathering crystals (including diamond) and rock fragments along the way. The high content of carbon dioxide dissolved in kimberlite magma becomes insoluble at lower pressures as the kimberlite ascends; carbon dioxide begins to form a separate gas phase at a depth of about 120 km, but most of the gas comes out at depths of 30 km and less. By the time the kimberlite breaks through the surface of the Earth, it is a suspension of liquid droplets and solid fragments in rapidly expanding gas. No lava flows representing liquid kimberlite that erupted on the surface have been recognized anywhere. This suggests that whenever kimberlite magma breaches the surface it does so explosively.

Kimberlites usually occur in regions of thick and stable continental crust, in southern Africa (including the Kimberley district), India, Siberia, Canada, Colorado-Wyoming, Venezuela, and Brazil. Most kimberlite outcrops appear on the surface as small, roughly circular areas less than 1 km in diameter; they are usually not well exposed because kimberlite weathers rapidly. In three dimensions, kimberlite bodies are dikes or, more commonly, downward-tapering cylinders (pipes). Those that have not been deeply eroded preserve their eruption craters, flaring out at the top of the pipe like the bell of a trumpet, filled with fragmental rock (kimberlite plus anything encountered on the way).

Because kimberlites usually do not crop out boldly, "indicator minerals" are important in prospecting for them. These are much more abundant than diamonds but occur with them in kimberlites and are stable under conditions of surface weathering, erosion, and deposition. They include pyrope garnet, ilmenite, and clinopyroxene, each with a distinctive composition. Using indicator minerals, one can trace debris from weathered kimberlite back to its source, even if that source has been buried.

During the violent injection of kimberlites into the upper crust, some detached fragments from the crust move downward and others from the lower crust and mantle move upward. Therefore kimberlite bodies are important scientifically because they contain fragments of rocks that were once above the present-day erosion surface as well as fragments of the Earth's mantle from 150 km deep or more. *See* IGNEOUS ROCKS.

Daniel S. Barker

Bibliography. A. J. A. Janse and P. A. Sheahan, Catalogue of world wide diamond and kimberlite occurrences: A selective and annotative approach, *J. Geochem. Explor.*, 53:73–111, 1995; R. H. Mitchell, *Kimberlites: Mineralogy, Geochemistry, and Petrology*, Plenum, New York, 1986; R. H. Mitchell, *Kimberlites, Orangeites, and Related Rocks*, Plenum, New York, 1995.

Kinematics

That branch of mechanics which deals with the motion of a system of material particles without reference to the forces which act on the system. Kinematics differs from dynamics in that the latter takes these forces into account. *See* DYNAMICS.

For a single particle moving in a straight line (rectilinear motion), the motion is prescribed when the position x of the particle is known as a function of the time. For uniformly accelerated motion, Eq. (1)

$$\frac{d^2x}{dt^2} = a \qquad (1)$$

holds, where a is the constant acceleration; x is defined as in Eq. (2), where x_0 and v_0 are the initial po-

$$x = {}^1/_2at^2 + v_0t + x_0 \qquad (2)$$

sition and velocity of the particle, respectively. *See* ACCELERATION; RECTILINEAR MOTION; VELOCITY.

Plane kinematics of a particle is concerned with the specification of the position of a particle moving in a plane by means of two independent variables, usually the rectangular cartesian coordinates x and y, but often the polar coordinates $r = (x^2 + y^2)^{1/2}$, $\theta = \tan^{-1}(y/x)$, or even other coordinates that may be especially convenient for a particular problem. Here the x and y components of velocity are dx/dt, dy/dt, and of acceleration d^2x/dt^2, d^2y/dt^2, respectively, but the components of velocity in the directions of increasing r and increasing θ are dr/dt and $r(d\theta/dt)$, and are therefore not so simply given in terms of the time derivatives of the coordinates.

The kinematics of a particle in space is concerned with the ways in which three independent coordinates may be chosen to specify the position of the particle at a given time, and with the relations between the first and second time derivatives of these coordinates and the components of velocity and acceleration of the particle. The motion is specified if three such coordinates are given as functions of the time, and the path of the particle in space is then obtained by eliminating the time between these equations.

For describing the simultaneous position in space of N particles, $3N$ coordinates are necessary, and the configuration of the system may be represented by a point in a 3*N*-dimensional space called the configuration space of the system. As the particles move, the corresponding representative point in configuration space traces out a curve called the trajectory of the system. If the particles do not move independently but are subject to *m* constraints, that is, if *m* relations between the 3*N* coordinates continue to hold throughout the motion, then it is possible to represent the trajectory in a space of $f = (3N - m)$ dimensions, *f* being the number of degrees of freedom of the system. For a rigid body, even though *N* and *m* are very large, $f = 6$, and the configuration at any time may be represented by six independent variables (usually three to represent the position of the center of mass and three angles to specify the orientation of the body). Kinematics describes the relations between the time derivatives of such angles and the components of the angular velocity of the body. *See* CAYLEY-KLEIN PARAMETERS; CONSTRAINT; DEGREE OF FREEDOM (MECHANICS); EULER ANGLES.

Among the coordinate systems studied in kinematics are those used by observers who are in relative motion. In nonrelativistic kinematics the time coordinate for each such observer is assumed to be the same, but in relativistic kinematics proper account must be taken of the fact that lengths and time intervals appear different to observers moving relative to each other. *See* RELATIVITY.

H. C. Corben; Benard Goodman

Kinetic methods of analysis

The measurement of reaction rates for the analytical determination of the initial concentrations of the species of interest taking part in chemical reactions. This technique can be used since, in most cases, the rates or velocities of chemical reactions are directly proportional to the concentrations of the species taking part in the reactions.

The rate of a chemical reaction is measured by experimentally following the concentration of some reactant or product involved in the reactions as a function of time as the mixture proceeds from a nonequilibrium to an equilibrium or static state (steady state). Thus, kinetic techniques of analysis have the inherent problem of the increased experimental difficulty of making measurements on a dynamic system, as time is now a variable which is not present in making measurements on an equilibrium system. However, kinetic methods often have advantages over equilibrium techniques in spite of the increased experimental difficulty. For example, the equilibrium differentiations or distinctions attainable for the reactions of very closely related compounds are often very small and not sufficiently separated to resolve the individual concentrations of a mixture without prior separation. But the kinetic differentiations or distinctions obtained when such compounds are reacted with a common reagent are often quite large and permit simultaneous analysis. *See* CHEMICAL EQUILIBRIUM.

A further advantage of kinetic methods is that they permit a larger number of chemical reactions to be used analytically. Many reactions, both inorganic and organic, are not sufficiently well behaved to be employed analytically by equilibrium or thermodynamic techniques. Many reactions attain equilibrium too slowly; side reactions occur as the reactions proceed to completion, or the reactions are not sufficiently quantitative (do not go to completion) to be applicable. However, a kinetic-based technique can often be employed in these cases simply by measuring the reaction rate of these reactions during the early or initial portion of the reaction period. Also, the measurement of the rates of catalyzed reactions generally is a considerably more sensitive analytical method for the determination of trace amounts of a large number of species than equilibrium methods. *See* CATALYSIS.

Kinetic methods of analysis can be divided into four basic categories: (1) methods employing uncatalyzed reactions, (2) methods employing catalyzed reactions, (3) enzyme-catalyzed reaction methods, and (4) methods for the simultaneous in-place determination of mixtures.

Uncatalyzed reactions. In order to determine the concentration of a single species in solution, the rate is measured of an irreversible uncatalyzed reaction of the type shown by reaction (1), where the single

$$A + R \xrightarrow{k_A} P \qquad (1)$$

species *A* reacts with the reagent *R* to form a product or products *P*; k_A is the rate constant for the reaction. In general, the types of reaction employed for kinetic analysis are second-order irreversible reactions. The rate of the reaction is most conveniently measured by following the rate of formation of the product *P* as a function of time (although the change in concentration of either *A* or *R* could also be followed). Any method of measuring concentrations, such as titrimetry, spectrophotometry, and polarography, can be employed. The rate of formation of the product, $d[P]_t/dt$, is given by differential equation (2).

$$\frac{d[P]_t}{dt} = k_A([R]_0 - [P]_t)([A]_0 - [P]_t) \qquad (2)$$

In this relation $[A]_0$ is the initial or original concentration of the species to be determined, $[R]_0$ is the initial concentration of the reagent (added in known concentration), and $[P]_t$ is the concentration of product formed at any time *t*; thus, the terms $([R_0 - [P]_t)$ and $([A_0 - [P]_t)$ equal the concentration of *A* and *R*, respectively, remaining at any time *t*, and k_A is the second-order rate constant for the reaction of *A* with *R*.

If the rate of reaction (1) is measured only during the initial portion of the reaction (during a time period chosen so that the reaction is only 2–3% complete), the concentration of product formed $[P]_t$ is small compared to $[A]_0$ and $[R]_0$. Thus, on rearrangement, Eq. (2) becomes Eq. (3). of Eq. (3) Integration

$$[A]_0 = \frac{d[P]_t/dt(\text{initial})}{k_A[R]_0} \qquad (3)$$

from the time interval $t = 0$ to t and for an initial concentration of $P = 0$ at $t = 0$ gives Eq. (4). Both fixed-

$$[A]_0 = \frac{[P]_t/t}{k_A[R]_0} \tag{4}$$

and variable-time methods may be used to calculate $[A]_0$ when employing Eq. (4).

When using the fixed-time method, the concentration of P formed at a chosen fixed time t_f is measured experimentally. As t_f and $[R]_0$ are thus kept constant for all measurements and k_A is a constant for all conditions, the term is the proportionality constant relating $[A]_0$ to $[P]_{tf}$, as shown in **Fig. 1**. For the variable-time method, with $[R]_0$ constant for all solutions, the time $t_{[P]f}$ necessary for the reaction to form a fixed amount of product $[P]_f$ is measured. As $[P]_f$ is now a constant, $[A]_0$ is related to $1/t_{[P]f}$ by the proportionately constant, $P_f/k_A[R]_0$, as shown in **Fig. 2**.

Catalyzed reactions. A catalyst may be broadly defined as an agent that alters the rate of a chemical reaction without shifting the equilibrium of the reaction. Although the catalyst undoubtedly enters into the reaction mechanism at a critical state, it does so in a cyclic manner and hence does not undergo a permanent change. Therefore a catalyst speeds up the rate of attainment of the equilibrium of a system, but it does not change the position of the equilibrium and it is not consumed during the reaction.

In many reactions involving a substance that acts as a catalyst, the concentration of the catalyst is directly proportional to the rate of the reaction. Thus the rates of these reactions can be employed for the analytical methods and are extremely sensitive, since the catalytic agents are not consumed but participate in the mechanism in a cyclic manner. The amount of the catalyst that can be determined employing catalyzed reactions is several orders of magnitude smaller than can be found by most direct equilibrium methods. Also, in many cases—enzyme-catalyzed (biological catalyst) reactions in particular—the catalyzed reaction of a reactant is extremely specific with respect to the chemical nature of the reactant. Such reactions, therefore, can be employed for the in-place analysis of a particular reactant in the presence of either a large excess or a large number of other species which would interfere with conventional equilibrium techniques, unless separation was performed prior to the analysis reaction.

The general mechanism for catalyzed reactions involves the combination of the catalyst C and the reactant, called the substrate S, to form a "complex" X, which then decomposes to form the product. Thus the catalyst is regenerated and combines once more with a substrate molecule. The general mechanism is given by reactions (5) and (6), where R is a reagent

$$C + S \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} X \tag{5}$$

$$X + R \xrightarrow{k_2} P + C \tag{6}$$

molecule (which in many catalyzed reactions is not needed) that reacts with the complex to give the product P, and k_1 and k_{-1} are the rate constants for the formation of the complex and for the conversion of the complex back to C and S, respectively; the equilibrium constant K for reaction (5) is equal to k_1/k_{-1}, and k_2 is the rate constant for the decomposition of X to form the product and C, as in reaction (6). In many catalyzed reactions the rate of the decomposition reaction (6) is slow compared to the rate of the conversion back to C and $S(k_2 \ll k_{-1})$. Thus, in this case, reaction (6) is the rate-determining step. Under these conditions, the rate of reaction of S in a catalyzed process is given by Eq. (7).

$$-\frac{d[S]_t}{dt} = \frac{k_2[C]_0[S]_0[R]_t}{[S]_0 + 1/K} \tag{7}$$

When the equilibrium constant K is relatively small, so that $[S]_0$ is small compared to $1/K$, and when R is in a nonlimiting excess (or nonexistent) such that its concentration does not change during the initial 2–3% of the reaction, Eq. (7) becomes Eq. (8),

$$-\frac{d[S]_t}{dt} = (K')[C]_0[S]_0 \tag{8}$$

where the constant K' equals $k_2K[R]_0$. Thus the initial

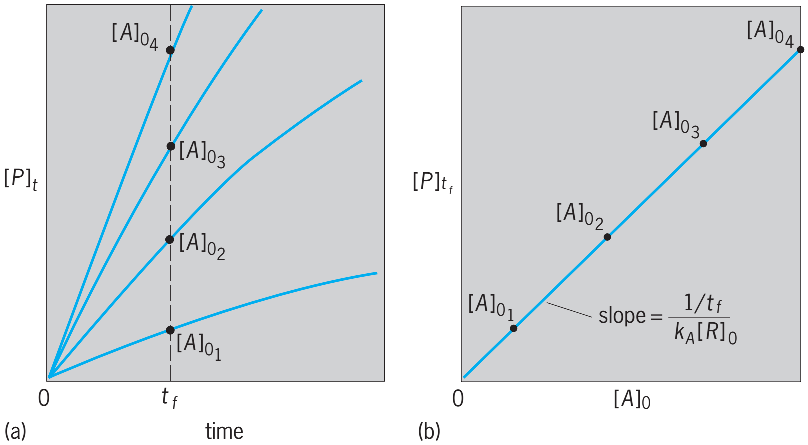


Fig. 1. **Fixed-time method for uncatalyzed reactions. (*a*) Variation of [*P*] as a function of time for different initial concentrations of [*A*]$_0$. (*b*) Variation of amount of *P* formed at fixed time t_f with initial concentration of *A*.**

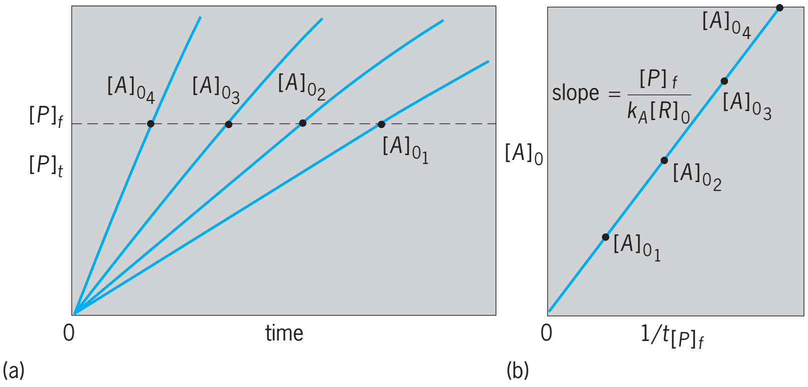


Fig. 2. **Variable-time method for uncatalyzed reactions. (*a*) Variation of [*P*] as a function of time for various initial concentrations of [*A*]$_0$. (*b*) Variation of $1/t_{[P]f}$ required to reach a fixed value of product formed $[P]_f$ with initial concentration of *A*.**

rate, $-(d[S]_t)/(dt)$, is directly proportional to the initial concentration of either the catalyst or the substrate, and the initial concentration of either can be determined. If a catalyst is to be determined, the substrate is added to each sample in nonlimiting excess (so that the concentration of S does not change during the initial 2–3% of the reaction). If a substrate is to be measured, the catalyst concentration added is not critical in any way because the catalyst is not consumed during the reaction. The choice of the catalyst concentration is dictated only by the range of initial rates that are conveniently measured. Either the fixed- or variable-time method of analysis can be employed by using catalyzed reactions.

Enzyme-catalyzed reactions. An enzyme is a biological catalyst, and this type of substance exhibits the ultimate in chemical selectivity. Enzyme-catalyzed reactions are used analytically for the determination of enzyme activities as well as selective substrate concentrations in complex biological and other mixtures. Both parameters are obviously of great importance in clinical diagnoses. The number of kinetic-based enzymatic determinations carried out daily in clinical laboratories is responsible for the fact that kinetic-based methods of analysis far outnumber those carried out by equilibrium-based methods or direct instrumental measurement.

In general, the simplified mechanism for enzyme-catalyzed reactions can be as shown in reaction (9),

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} ES^{+} \xrightarrow{k_2} P + E \qquad (9)$$

where E = enzyme, S = substrate, ES^+ = the activated addition complex, and P = products. Equation (10) is the differential rate equation for reaction (9).

$$-\frac{d[S]_t}{d_t} = \frac{d[P]_t}{d_t} = \frac{k_2[E]_0[S]_t}{K_m + [S]_t} \qquad (10)$$

$$K_m = \frac{k_{-1} + k_2}{k_1}$$

As in catalyzed reactions, experimental conditions can be adjusted for either substrate analysis or enzyme analysis employing the initial portion of the reaction (only 2–3% complete). Again, either the fixed- or variable-time method of analysis can be employed by using enzyme-catalyzed reactions. *See* CLINICAL PATHOLOGY; ENZYME.

Simultaneous determinations. Often the thermodynamic properties of closely related species are such that conventional equilibrium analytical techniques are unable to resolve the analytical concentrations of the components of the mixture without prior separation. However, in general the rates of reaction of such species in a mixture are sufficiently different to enable one to determine the initial concentration of each species without resorting to separation, which is a considerable saving of time and labor. Such techniques are called differential kinetic analysis methods. There are a large number of differential kinetic methods applicable to both first- and second-order reactions. Each of these methods has special conditions under which it is most applicable. The principles of these methods are complex, and an explanation of each is beyond the scope of this section. However, the detailed explanation of the logarithmic extrapolation method given below illustrates the general concepts of differential kinetic methods.

Consider two competing second-order irreversible reactions of the type shown by (11) and (12),

$$A + R \xrightarrow{k_A} P \qquad (11)$$

$$B + R \xrightarrow{k_B} P \qquad (12)$$

where k_A and k_B are the second-order rate constant for the reaction of A and B with R, respectively. If the concentration of R is 50 to 100 times greater than the sum of the concentration of A plus B (the components of interest in the analysis), the reaction is pseudo first-order because $[R]_t$ is constant and is equal to the initial concentration $[R]_0$. The differential rate expressions for the reactions given by (11) and (12) then become Eqs. (13) and (14).

$$-\frac{d[A]_t}{dt} = k_A[R]_0[A]_t = k'_A[A]_t \qquad (13)$$

$$-\frac{d[B]_t}{dt} = k_B[R]_0[B]_t = k'_B[B]_t \qquad (14)$$

The sum of the concentrations of A and B which react competitively to form a common product P is given at any time t by the rate expression, as shown by Eq. (15).

$$\frac{d[P]_t}{dt} = -\left[\frac{d[A]_t}{dt} + \frac{d[B]_t}{dt}\right]$$
$$= k'_A[A]_t + k'_B[B]_t \qquad (15)$$

On integration of Eq. (15) between the limits $t = 0$ and $t = \infty$ Eq. (16) is formed. In the case where

$$[P]_\infty - [P]_t = [A]_t + [B]_t$$
$$= [A]_0 e^{-k'_A t} + [B]_0 e^{-k'_B t} \qquad (16)$$

the rate of reaction of component A is larger than that of B, the term $[A]_0 e^{-k'_A t}$ eventually becomes very small compared to $[B]_0 e^{-k'_B t}$ at some time t after A has reacted essentially to completion ($[A_t] \cong 0$) and can be considered negligible. Thus, by taking the logarithm of both sides of Eq. (16), Eq. (17), is

$$\ln([A]_t + [B]_t) = \ln([P]_\infty - [P]_t)$$
$$= -k'_B t + \ln[B]_0 \qquad (17)$$

obtained, which predicts that a plot $\ln([A]_t + [B]_t)$ or $\ln([P]_\infty - [P]_t)$ versus time t will yield a straight line with a slope of $-k'_B$ and an intercept (at $t = 0$) equal to $\ln[B]_0$. The value of $[A]_0$ is then obtained by subtracting $[B]_0$ from the total initial concentration of the mixture, $[A]_0 + [B]_0$, which must be determined either by independent methods or from $[P]_\infty$. A typical reaction rate curve of this type is illustrated in **Fig. 3**. *See* CHAIN REACTION (CHEMISTRY); CHEMICAL DYNAMICS; ELECTROCHEMICAL TECHNIQUES;

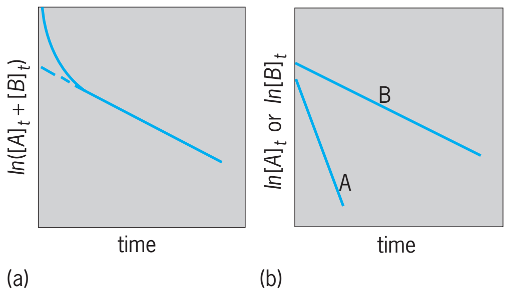


Fig. 3. **Logarithmic extrapolation method for a mixture of species *A* and *B* reacting by first-order kinetics. (*a*) Rate data obtained for the mixture. (*b*) Rate data obtained for each component separated.**

INHIBITOR (CHEMISTRY); KINETIC THEORY OF MATTER. Harry B. Mark, Jr.

Bibliography. S. W. Benson, *Foundations of Chemical Kinetics*, 1960, reprint 1982; W. Drenth and H. Kwart, *Kinetics Applied to Organic Reactions* 1980; S. Kotrly and S. Ladislav, *Handbook of Chemical Equilibria in Analytical Chemistry,* 1985; H. B. Mark, Jr., and G. A. Rechnitz, *Kinetics in Analytical Chemistry*, 1968; H. A. Mottola, *Kinetic Aspects of Analytical Chemistry*, 1988; D. Pérez-Bendito and M. Silva, *Kinetic Methods in Analytical Chemistry*, 1988; R. W. Ramette, *Chemical Equilibrium and Analysis*, 1981.

Kinetic theory of matter

A theory which states that the particles of matter in all states of aggregation are in vigorous motion. In computations involving kinetic theory, the methods of statistical mechanics are applied to specific physical systems. The atomistic or molecular structure of the system involved is assumed, and the system is then described in terms of appropriate distribution functions. The main purpose of kinetic theory is to deduce, from the statistical description, results valid for the whole system. The distinction between kinetic theory and statistical mechanics is thus of necessity arbitrary and vague. Historically, kinetic theory is the oldest statistical discipline. Today a kinetic calculation refers to any calculation in which probability methods, models, or distribution functions are involved.

For information which is related to and supplements the present article *see* BOLTZMANN STATISTICS; BOLTZMANN TRANSPORT EQUATION; QUANTUM STATISTICS; STATISTICAL MECHANICS.

Classes of problems. Kinetic calculations are not restricted to gases, but occur in chemical problems, solid-state problems, and problems in radiation theory. Even though the general procedures in these different areas are similar, there are a sufficient number of important differences to make a general classification useful.

Classical ideal equilibrium problems. In these, there are no interactions between the constituents of the system. The system is in equilibrium, and the mechanical laws governing the system are classical. The basic information is contained in the Boltzmann distribution f (also called Maxwell or Maxwell-Boltzmann distribution) which gives the number of particles in a given momentum and positional range ($d^3x = dx\,dy\,dz$, $d^3\upsilon = d\upsilon_x\,d\upsilon_y\,d\upsilon_z$, where x, y, and z are coordinates of position, and υ_x, υ_y, and υ_z are coordinates of velocity). In Eq. (1) ϵ is the energy,

$$f(xyz, \upsilon_x\upsilon_y\upsilon_z) = Ae^{-\beta\epsilon} \tag{1}$$

$\beta = 1/kT$ (where k is the Boltzmann constant and T is the absolute temperature), and A is a constant determined from Eq. (2). The calculations of gas pressure,

$$\int\int d^3x\, d^3\upsilon\, f = N \quad \text{total number of particles} \tag{2}$$

specific heat, and the classical equipartition theorem are all based on these relations.

Classical ideal nonequilibrium problems. Many important physical properties refer not to equilibrium but to nonequilibrium states. Phenomena such as thermal conductivity, viscosity, and electrical conductivity all require a discussion starting from the Boltzmann transport equation, Eq. (3*a*), where $\mathbf{X}$ is the force

$$\frac{\partial f}{\partial t} + \mathbf{v}\cdot\nabla f + \mathbf{X}\cdot\nabla_{\upsilon} f = \left(\frac{\partial f}{\partial t}\right)_{\text{coll}} \tag{3a}$$

$$Z_{cl} = e^{-\Psi/kT} = \frac{1}{N!\lambda^{3N}}\int\cdots\int d^3x_1 \cdots d^3x_N \exp[-(1/kT)U_{(x_1,\ldots,z_N)}] \tag{3b}$$

per unit mass, ∇ is the gradient on the space-dependence, ∇_υ is the gradient on the velocity-dependence, and $(\partial f/\partial t)_{\text{coll}}$ is the change due to collisions. If one deals with states that are near equilibrium, the exact Boltzmann equation need not be solved; then it is sufficient to describe the nonstationary situation as a small perturbation superimposed on an equilibrium state. Even though the rigorous discussion of the nonequilibrium processes is difficult, appeal to simple physical pictures frequently leads to quite tractable expressions in terms of the equilibrium distribution function. Of special importance is the example of the so-called electron gas in a metal. The kinetic treatment of this system forms the basis for the classical (Lorentz) conductivity theory. For states far from equilibrium, no general simple theory exists. *See* FREE-ELECTRON THEORY OF METALS.

Classical nonideal equilibrium theory. The basic classical procedure for arbitrary systems (systems with interactions taken into account) that allows the calculation of macroscopic entities is that using the partition function, as shown in Eq. (3*b*), where $\lambda = h/\sqrt{2\pi mkT}$. Here Ψ is the thermodynamic free energy, and h is Planck's constant. Although Eq. (3*b*) is written so that it may be applied to gases, the partition function may also be written for classical spin systems, such as ferromagnetic and paramagnetic solids. The mathematical problems of evaluating the integrals or the sums are difficult. Equation (3*b*), with appropriate modifications, is the starting point for all these considerations. *See* FREE ENERGY.

Classical nonideal nonequilibrium theory. This is the most general situation that classical statistics can describe. In general, very little is known about such systems. The Liouville equation applies, and has been used as a starting point for these studies, but no spectacular results have yet been obtained. For studies of the liquid state, however, the results have been quite promising.

Quantum problems. There are quantum counterparts to the classifications just described. In a quantum treatment a distribution function is also used for an ideal system in equilibrium to describe its general properties. For systems of particles which must be described by symmetrical wave functions, such as helium atoms and photons, one has the Bose distribution, Eq. (4), where $\beta = 1/kT$, and A is determined by Eq. (2).

$$f(v_x v_y v_z) = \frac{1}{(1/A)e^{\beta\epsilon} - 1} \tag{4}$$

See BOSE-EINSTEIN STATISTICS.

For systems of particles which must be described by antisymmetrical wave functions, such as electrons, protons, and neutrons, one has the Fermi distribution, Eq. (5). Use of these functions in calculations is actually not very different from the use of the classical distribution function, but the results are quite different, as are the analytical details. As in the classical case, the treatment of the nonequilibrium state can be reduced to a treatment involving the equilibrium distribution only. The application to electrons as an (ideal) Fermi-Dirac gas in a metal is the basis of the Sommerfeld theory of metals. *See* FERMI-DIRAC STATISTICS.

$$f(v_x v_y v_z) = \frac{1}{(1/A)e^{\beta\epsilon} + 1} \tag{5}$$

In quantum theory, nonideal systems in equilibrium are described in terms of the quantum partition function, Eq. (6). Here E_n indicates the energy levels of the system, and g_n indicates the weights of these levels. If one defines a Slater sum by Eq. (7), where $U_n(x_1, \ldots, x_N)$ is the wave function of the state n, Z_g may be written as an integral similar to Z, as in Eq. (8). For the applications, the energy levels and the wave functions must be known. In the evaluation of S, given by Eq. (7), the symmetry character of the wave functions must be explicitly introduced. It is sometimes easier to use the grand partition function, Eq. (9). Here μ is the chemical potential, and $E_{N,n}$ is the nth level of a system having N particles. The current theories of quantum statistics, as for example the hard-sphere Bose gas, are for the most part concerned with questions in this area.

$$Z_g = \sum_n g_n e^{-E_n/kT} = e^{-\Psi/kT} \tag{6}$$

$$S(x_1, \ldots, x_N) = \sum_n e^{-E_n/kT} \, |U_n(x_1, \ldots, x_n)|^2 \tag{7}$$

$$Z_g = \int \cdots \int d^3x_1 \cdots d^3x_N S(x_1, \ldots, x_N) \tag{8}$$

$$Z_{q\cdot m\cdot gr} = \sum_N \sum_n e^{(\mu N - EN,n)/kT} \tag{9}$$

The only technique now available is that of the density matrix. It is possible to express certain entities which characterize transport properties, such as the conductivity tensor, in terms of the unperturbed stationary density matrix. A complete discussion of the validity of the approximations is still lacking. In addition, once the conductivity tensor is obtained in terms of the density matrix of the stationary (but still interacting) system, a problem of the same order of difficulty as the evaluation of the quantum mechanical partition function remains, if explicit expressions for these quantities in terms of the forces between atoms are desired.

Classical examples. Kinetic theory gave the first insight into many of the phenomena that take place in gases as well as in metals, where the free (conduction) electrons can be considered as an ideal gas of electrons. The following examples illustrate some of the more fundamental calculations that have been made.

Ideal-gas pressure. A classical ideal gas is described by the Boltzmann distribution of Eq. (1). The constant A is given by Eq. (10), where m is the mass of an individual molecule and V is the total volume of the gas. A gas exerts a force on the wall by virtue of the fact that the molecules are reflected by it. The component of momentum normal to the wall changes its sign as a consequence of this collision. Hence if the normal velocity of the molecule is v_n, the momentum given off to the wall is $2mv_n$. To calculate the total force on a wall, one needs the total momentum transferred to the wall per unit time. Let dS be a small section of the wall, and call θ the angle made by the molecule's velocity vector with the normal to dS (**Fig. 1**). From Eq. (1), the number of molecules per unit volume that have a speed $c = (v_x^2 + v_y^2 + v_z^2)^{1/2}$ and whose velocity vector makes an angle θ with a given axis (call these c, θ molecules) is given by Eq. (11*a*). The number of such molecules that in time dt will collide with dS (assuming spatial homogeneity) is given by notation (11*b*). From this information an interesting side result may be calculated, namely, that the number of all collisions with a unit area of the wall per second may be written as expression (12*a*). If one introduces the average speed $\bar{c}$, defined in Eq. (12*b*),

$$A = \frac{N}{V}\left(\frac{m\beta}{2\pi}\right)^{3/2} \tag{10}$$

$$f(c, \theta) = 2\pi A e^{(-1/2)\beta mc^2} c^2 \sin\theta \, d\theta dc \tag{11a}$$

$$dS \, c\cos\theta \, dt \, f(c, \theta) \tag{11b}$$

$$2\pi A \int_0^{\pi/2} d\theta \;\sin\theta\cos\theta \int_0^{\infty} dc \; c^3 e^{-(1/2)\beta mc^2} \tag{12a}$$

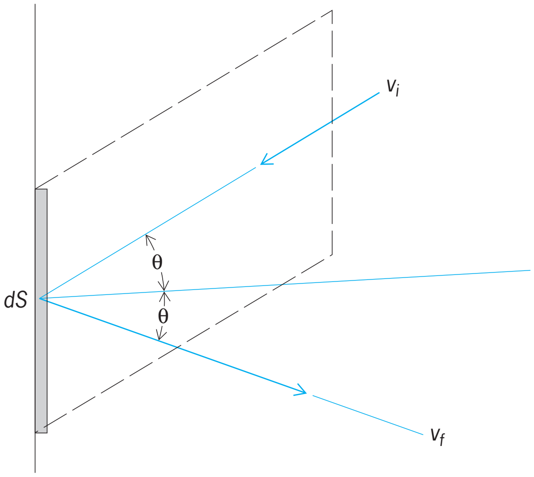


Fig. 1. Change of momentum of gas molecule as it strikes a wall; v_i, initial velocity; v_f, final velocity.

$$\bar{c} = \frac{1}{N} \int \int d^3x \, d^3v \, c f \qquad (12b)$$

the total number of collisions may be written as $^1/_4 n\bar{c}$. Here $n = N/V$, the number density. This relation is of importance in calculating the efflux of gases through orifices. Each (c, θ) molecule contributes a momentum of $2mc \cos\theta$ to the wall per collision. Since the pressure p is given by the force per unit area, or the total momentum transferred per unit time per unit area, one obtains for p Eqs. (13) and (14). Since it

$$p = \int_0^\infty \int_0^{\pi/2} 2mc \, \cos\theta \cdot 2\pi A c \cos\theta \, e^{-(1/2)\beta mc^2} c^2 \sin\theta \, d\theta \, dc \qquad (13)$$

$$p = {}^1/_3 N m\overline{c^2} = (N/V)\beta^{-1} \qquad (14)$$

is known experimentally that $pV = NkT$, where k is the Boltzmann constant, $\beta = 1/kT$ is now identified. This pressure calculation is a typical example of a kinetic theory calculation. *See* GAS.

Equipartition theorem and specific heat. In a classical ideal gas, the average kinetic energy associated with any translational degree of freedom is expressed by Eq. (15). It is quite straight forward to show that

$$^1/_2 m\overline{v_x^2} = {}^1/_2 m\overline{v_y^2} = {}^1/_2 m\overline{v_z^2} = \frac{1}{2\beta} = {}^1/_2 kT \qquad (15)$$

the average energy associated with any degree of freedom, which occurs as a quadratic term in the expression for the mechanical energy, is given by $^1/_2 kT$. Formally, the result one proves is that, for a Boltzmann distribution written in terms of momenta and generalized coordinates p_i and q_i, one has Eq. (16), which is called the equipartition theorem.

$$\overline{q_i \frac{\partial \epsilon}{\partial q_i}} = \overline{p_i \frac{\partial \epsilon}{\partial p_i}} = kT \qquad (16)$$

From the equipartition theorem the specific heat C, defined by Eq. (17), may be immediately obtained.

$$C = \frac{\partial \overline{E}}{\partial T} \qquad (17)$$

If one has just the three translational degrees of freedom, the average energy per degree of freedom per molecule is $^1/_2 kT$; hence the average energy $\overline{E} = {}^3/_2 NkT$ and $C = {}^3/_2 R$. The specific heat is thus constant and independent of T. In the case of a diatomic molecule, one usually has three translational and two rotational degrees of freedom. Therefore $\overline{E} = {}^5/_2 NkT$, and $C = {}^5/_2 Nk \equiv {}^5/_2 R$. Suppose there are N atoms in a solid, each bound by elastic forces to a center. In that case, the mechanical energy will be expressed by Eq. (18), where k_F is the elastic force instant.

$$\epsilon = \frac{1}{2m}\left(p_x^2 + p_y^2 + p_z^2\right) + {}^1/_2 k_F^2 (x^2 + y^2 + z^2) \qquad (18)$$

According to the equipartition theorem, these six terms in ϵ will give a specific heat $C = 3Nk$, the Dulong and Petit value for a monatomic solid. *See* SPECIFIC HEAT OF SOLIDS.

The equipartition theorem is especially useful when a gas is at a temperature so high that it becomes necessary to use relativistic mechanics to describe the system properly. This happens at temperatures for which the thermal energy kT is of the same order as mc^2, where c denotes the speed of light. The appropriate relation between energy and momentum is then given by Eq. (19*a*). The equipartition relation (16) may then be written as Eq. (19*b*). From this relation one can obtain the relativistic correction to the average energy of a gas, as shown in Eq. (19*c*).

$$\epsilon = c\left[\left(p_x^2 + p_y^2 + p_z^2\right) + m^2c^2\right]^{1/2} \qquad (19a)$$

$$\overline{\frac{c^2 p_x^2}{\epsilon}} = kT \qquad (19b)$$

$$\overline{E} = {}^3/_2 NkT \left(1 + \frac{5}{4}\frac{kT}{mc^2} + \cdots\right) \qquad (19c)$$

See RELATIVISTIC MECHANICS.

Electrical conductivity of metals. Experimentally, a proportionality is observed between the applied electric field and the current produced in a metallic conductor. Kinetic procedures give an explanation of this general connection. Consider the electrons in a metal as an ideal gas of electrons of mass m and charge e. In equilibrium, the electrons are described by the Boltzmann distribution, Eq. (1), or the Fermi-Dirac distribution, Eq. (5). Call this distribution f. Because f depends on $v_x^2 + v_y^2 + v_z^2$ only, it is clear that $\bar{v}_x = 0$; no net current can flow in an equilibrium state. The application of an electric field therefore results in the destruction of the spherical symmetry in the velocities. This effect is described by the Boltzmann transport equation, Eq. (20).

$$\frac{\partial f}{\partial t} + \mathbf{v} \cdot \nabla f + \mathbf{X} \cdot \nabla_v f = \left(\frac{\partial f}{\partial t}\right)_{\text{coll}} \qquad (20)$$

The collision term causes considerable difficulty. In many problems one is justified in introducing a relaxation time τ which may depend on x and v and which is defined by Eq. (21). Here f_0 is the distribution function at thermal equilibrium.

$$\left(\frac{\partial f}{\partial t}\right)_{\mathrm{coll}} = -\frac{f - f_0}{\tau} \qquad (21)$$

The introduction of such a relaxation time presupposes that a nonequilibrium state will decay exponentially into an equilibrium state as a consequence of the action of the collisions. This is undoubtedly a good approximation near equilibrium. *See* RELAXATION TIME OF ELECTRONS.

Assume the existence of an electric field in the x direction which distorts the initial distribution a small amount. By using Eq. (21), Eq. (20) may be written as Eq. (22). Here E_e is the external electric field.

$$\frac{eE_e}{m}\frac{\partial f}{\partial v_x} + v_x\frac{\partial f}{\partial x} = -\frac{f - f_0}{\tau} \qquad (22)$$

If it is now assumed that $(f - f_0)/f_0 \ll 1$, with f very near an equilibrium state, so that quadratic terms may be neglected, and that the parameters A and β in f_0 are independent of x, then an explicit expression for f can be obtained, as shown in Eq. (23).

$$f = f_0 - \tau e E_e v_x \frac{\partial f_0}{\partial \epsilon} \qquad (23)$$

The electric current in the x direction is always given by Eq. (24).

$$j_x = \int e v_x f\, d^3v = -\tau e^2 E_e \int v_x^2 \frac{\partial F_0}{\partial \epsilon} d^3v \qquad (24)$$

The first term in Eq. (23) does not contribute to the current, as has already been pointed out. This is an example of a formal result in transport theory. The result for j_x is in a form in which only a knowledge of the equilibrium distribution is required to obtain an explicit answer.

In the case of Boltzmann statistics, use of Eqs. (1), (2), (10), and (24) gives Eq. (25*a*). The conductivity

$$j_x = \frac{Ne^2\tau}{m}E_e \qquad (25a)$$

$$f_0 = \frac{1}{e^{\beta(\epsilon-\mu)} + 1} \qquad (25b)$$

$\sigma = Ne^2\tau/m$ cannot be compared with experiment unless the relaxation time τ is known. In the Fermi case the evaluation of Eq. (24) is facilitated by observing that $\partial f_0/\partial\epsilon$ has a δ-function character. Call $A = e^{\beta\mu}$, so that Eq. (5) reads as Eq. (25*b*).

It is easy to show from Eq. (25*b*) that expression (26) holds for sufficiently low temperatures. This allows the immediate calculation of Eq. (24) also for the case in which the relaxation time depends on the velocity.

$$\frac{\partial}{\partial\epsilon}\frac{1}{e^{\beta(\epsilon-\mu)} + 1} \cong -\delta(\epsilon - \mu) \qquad (26)$$

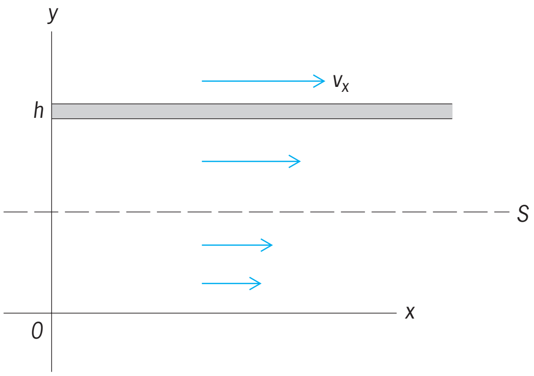


Fig. 2. Explanation of viscosity. The plate at $y = h$ is moving with velocity v_x, and the plate at $y = 0$ is stationary. The horizontal component of velocity of the gas molecules varies from 0 at $y = 0$ to v_x at $y = h$. Momentum therefore must be transferred across surface S by the vertical component of velocity.

Viscosity and mean free path. One of the early successes of kinetic theory was the explanation of the viscosity of a gas. Strictly speaking, this is again a transport property, and as such it should be obtained from the Boltzmann transport equation, Eq. (20). It is possible, however, to give an elementary discussion. Consider a gas that is contained between two walls or plates, the lower one ($y = 0$) at rest and the upper one constrained to move with a given velocity in the x direction (**Fig. 2**). A force is necessary to maintain the constant velocity of the plate. This force is given by Eq. (27). Here dv_x/dy is the velocity gradient, and

$$\mathbf{X}_x = \eta A \frac{dv_x}{dy} \qquad (27)$$

X is the viscous force on the area A, which is perpendicular to the y axis ($\mathbf{X}_x/A$ is sometimes called the shear stress). Equation (27) defines η, the viscosity coefficient. The physical reason for this force stems from the fact that molecules above the surface S have a greater flow velocity than those below this surface. (This will certainly be true on the average.) Thus molecules crossing from above to below will carry a larger amount of momentum in the positive x direction than those crossing from below S upward. Hence the net effect is a transport of momentum in the x direction across the surface. By Newton's second law, this will yield a force. The computation can be carried out in this manner. Consider an area in the xz plane. The number of molecules passing through per second is $fv_y dS$, where f is the distribution function. The amount of momentum transported in the x direction is (per collision) $fv_y dS \cdot mv_x$.

The force per unit area is the sum of these terms, given by expression (28). For the evaluation of this sum, the notion of mean free path is useful. The mean free path is the average distance traveled by a molecule between collisions, and is usually designated by λ. To investigate this entity, imagine each molecule to be a hard sphere with radius a. If a molecule moves with average speed $\bar{c}$, it sweeps out a volume $4\pi a^2\bar{c}t$ in time t. If there are $n = N/V$ molecules per unit volume, the number of collisions per second is given by the collision frequency z in Eq. (29).

$$fv_y v_x m \qquad (28)$$

$$z = n4\pi a^2\bar{c} \qquad (29)$$

For a typical gas (oxygen) under standard conditions $n = 3 \times 10^{25}$, $\bar{c} = 4.5 \times 10^4$ cm/s, and $a \cong 1.8 \times 10^{-12}$ cm. Hence, numerically, $z = 5.5 \times 10^9$ collisions/s. The average distance between collisions, that is, the mean free path, is given by Eq. (30*a*). Numerically, $\lambda \cong 8 \times 10^{-6}$ cm. This discussion is, of course, exceedingly crude. Making the calculation on the basis of a Boltzmann distribution gives Eq. (30*b*).

$$\lambda = \frac{\bar{c}t}{n(4\pi a^2)\bar{c}t} = \frac{1}{n(4\pi a^2)} \tag{30a}$$

$$\lambda = \frac{1}{\sqrt{2}n(4\pi a^2)} \tag{30b}$$

Using similar methods, it may be shown that the distribution of the mean free paths (that is, the number of molecules whose mean free path lies between x and $x + dx$) is given by Eq. (31).

$$dN = \frac{N_0}{\lambda} e^{-x/\lambda}\, dx \tag{31}$$

There is an interesting connection between the mean free path and the relaxation time introduced previously. It should be stressed, however, that this connection follows more from a qualitative discussion than from a rigorous calculation. One would guess that Eq. (32) is valid. This means that a relaxation time describes the decay from a state so near an equilibrium state that, when the molecules have traveled (on the average) one mean free path, the equilibrium is reestablished. Stated differently, the nonequilibrium state is, on the average, one collision per molecule removed from the equilibrium state.

$$\tau = \frac{\lambda}{\bar{c}} \tag{32}$$

Expression (28) may now be evaluated in terms of the mean free path as shown in Eq. (33*a*). From this, the viscosity coefficient follows directly, as shown in Eq. (33*b*). Introducing the mean free path, as given by Eq. (30*a*), one sees that Eq. (33*c*) holds. The remarkable result is that the viscosity is indeed independent of the pressure. Since Eq. (34) holds, the viscosity depends on the temperature but not on the pressure. This somewhat unintuitive result was one of the first triumphs of kinetic theory. Both aspects of Eq. (33*c*) are in good agreement with experiment; for not too low densities η is indeed independent of the pressure; η is also proportional to the square root of the temperature. *See* VISCOSITY. Max Dresden

$$\text{Force per unit area} = {}^1/_3 nm\lambda\bar{c}\frac{dv_x}{dy} \tag{33a}$$

$$\eta = {}^1/_3 nm\bar{c}\lambda \tag{33b}$$

$$\eta = \frac{1}{3}\frac{m\bar{c}}{4\pi a^2} \tag{33c}$$

$$\bar{c} = \sqrt{8kT/\pi m} \tag{34}$$

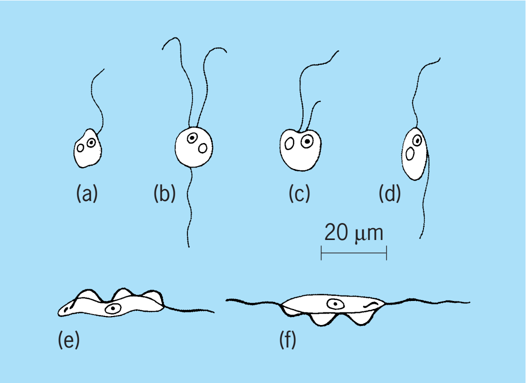
Bibliography. R. L. Liboff, *Kinetic Theory: Classical, Quantum and Relativistic Description*, 2d ed., 1998; P. C. Reidi, *Thermal Physics: An Introduction to Thermodynamics, Statistical Mechanics, and Kinetic Theory*, 2d ed., 1988; P. P. Schram, *Kinetic Theory of Gases and Plasmas*, 1991; L. C. Woods, *An Introduction to the Kinetic Theory of Gases and Magnetoplasmas*, 1993.

Kinetoplastida

An order of the class Zoomastigophorea in the phylum Protozoa, also known as Protomastigida, containing a heterogeneous group of colorless flagellates possessing one or two flagella in some stage of their life cycle. These small organisms (5–89 micrometers in length) typically have pliable bodies. Some species are holozoic and ingest solid particles, while others are saprozoic and obtain their nutrition by absorption. Their life cycles are usually simple but some species have two or more recognizably distinct stages. The species may be either free-living or parasites of vertebrates, invertebrates, and plants. Reproduction is by longitudinal fission, although multiple fission occurs in some species. Several are important disease-causing parasites of humans and domestic and wild animals.

Taxonomy. The order Kinetoplastida includes all of the Protozoa with only one or two flagella that are not markedly ameboid, do not contain chloroplasts, and are not closely related to chloroplast-bearing flagellates. Organisms that possess one or two flagella and are decidedly ameboid belong to the family Mastigamoebidae in the order Rhizomastigida. Those which are colorless but considered to be close relatives of protozoans possessing chloroplasts belong to the family Chlamydomonadidae of the order Volvocida. *See* VOLVOCIDA.

There is disagreement on the division of the order into families. However, the five or more families can be divided into two general groups. The first group contains simple organisms with no distinctive features save one or two flagella of equal or



Representative genera of families of order Kinetoplastida. (*a*) *Oikomonas* (family Oikomonadidae), one anterior flagellum. (*b*) *Amphimonas* (family Amphimonadidae), two equally long anterior flagella. (*c*) *Monas* (family Monadidae), two unequally long anterior flagella. (*d*) *Bodo* (family Bodonidae), two unequally long flagella, one of them trailing. (*e*) *Trypanosoma* (family Trypanosomatidae), one flagellum with undulating membrane. (*f*) *Cryptobia* (family Cryptobiidae), two flagella, one free and one with undulating membrane.

unequal length. This includes the families Oikomonadidae, Amphimonadidae, Monadidae, and Bodonidae (**illus.** *a–d*). The second group contains organisms which have an undulating membrane in addition to one or two flagella. The families included in this group are Trypanosomatidae, and Cryptobiidae (illus. *e* and *f*). A third group, including organisms possessing a peculiar collar surrounding a single flagellum, once assigned to this order, now makes up the order Choanoflagellida. *See* CHOANOFLAGELLIDA.

Trypanosomatidae. The most important family of the Kinetoplastida is the Trypanosomatidae, since it includes several species that infect humans and domestic animals with serious diseases, such as African sleeping sickness. The organisms in this family are polymorphic, changing their form in various stages of their development. Their life cycles may involve one or two hosts (invertebrate, vertebrate, or plant). The trypanosome form possesses a single flagellum and an undulating membrane extending the full length of the body. The other related forms are simpler in that they lack one or both of these structures. *See* CILIA AND FLAGELLA; MASTIGOPHORA; PORIFERA; PROTOZOA; TRYPANOSOMATIDAE. M. M. Brooke

Bibliography. S. H. Hutner and A. Lwoff (eds.), *Biochemistry and Physiology of Protozoa,* 3 vols., 1961–1964; T. L. Jahn and F. F. Jahn, *How to Know Protozoa,* 1949; R. R. Kudo, *Protozoology,* 5th ed., 1977; A. Westfall, *Protozoa*, 1976.

Kinorhyncha

A phylum of free-living marine invertebrates less than 1 mm long. They are segmented and lack internal ciliation (see **illustration**). Kinorhyncha, formerly called Echinoderida, are sometimes included in the rejected superphylum Aschelminthes, since their fluid-filled body cavity is probably a pseudocoel, or false coelom. Kinorhynchs are benthonic, generally dwelling in mud or sand from intertidal to deep-sea habitats. Two orders are generally recognized, Cyclorhagida and Homalorhagida.

The body is covered by a transparent cuticle secreted by an underlying epidermis. The cuticle is molted only in the process of juvenile growth. Three body regions are recognized: a head segment, a neck segment, and an 11-segment trunk. The head is completely retractable. When everted, the head extends its five to seven circles of recurved spines called scalids.

The neck consists of plates called placids which function to close the anterior opening of the trunk when the head is retracted. The Cyclorhagida have 14–16 placids arranged around the circular opening of the trunk; but in one suborder, the Conchorhagae, these placids function with the two lateral clam-shell portions of the trunk segment which constitutes the primary closing mechanism. In the Homalorhagida four to eight placids function, with the three ventral plates of the first trunk segment, like a trapdoor—the anterior margin of which is drawn up against a vaulted dorsal plate to close the triangular-shaped opening into which the head is withdrawn.

Most kinorhynchs have a pair of adhesive tubes on the ventral surface of the third or fourth segment. Additional adhesive tubes may occur near some of the lateral spines of more posterior trunk segments.

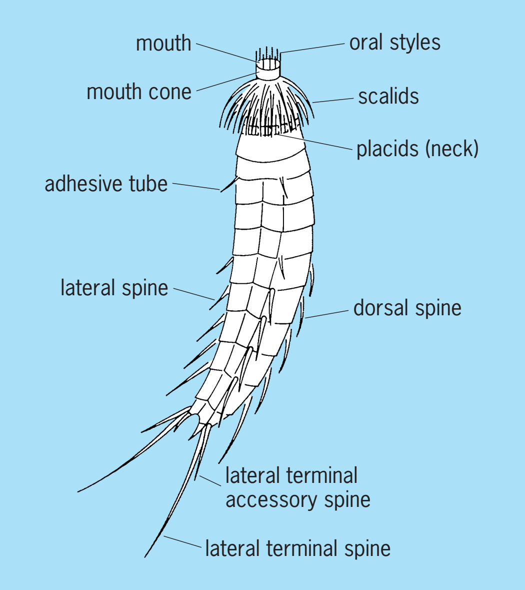
Circular musculature is present in the head, neck, and pharynx. Trunk musculature is striated and consists of segmentally arranged longitudinal, oblique, and dorsoventral muscles. Anteriorly, the longitudinal muscles form the retractor muscles of the head. Eversion of the head is accomplished by simultaneously relaxing the retractor muscles while contracting paired dorsal ventral muscles, which increases the pressure of the body fluid and everts the head. Retraction of the head is the reverse of this process.

Locomotion is a consequence of repeated eversion and retraction of the head. As the head is everted, mud or sand particles are gripped by the scalids, and the animal is pulled forward.

The nervous system consists of a 10-lobed brain which encircles the anterior end of the pharynx, a poorly developed double ventral nerve cord with segmentally arranged groups of ganglionic cells, and eight other longitudinal nerves situated dorsally, subdorsally, laterally, and lateroventrally.

The mouth is located terminally on the mouth cone, which protrudes from the center of the head. A series of anteriorly directed spines surround the mouth opening. There is a short buccal cavity followed by a muscular pharynx similar to that of the Nematoda and Gastrotricha, except that it is lined with a syncytial epithelium which overlies the cuticle. The pharynx is followed by a short esophagus and stomach-intestine, which is a straight tube with a network of muscle fibers. The end gut is short, lined with cuticle, and separated from the stomach-intestine by a sphincter. The anus is terminal.

A single pair of protonephridial excretory organs



Echinoderes **sp., a cyclorhagid kinorhynch (ventrolateral view).**

begins in the tenth segment and opens laterodorsally in the eleventh segment. These organs are modified solenocytes, having either a single long driving flagellum as in the Cyclorhagida, or a long flagellum and several shorter driving flagella along the canal in the Homalorhagida.

The Kinorhyncha are dioecious. In each sex there is one pair of gonads. The genital pores are located on the thirteenth segment. The ovary contains both ova and nutritive cells. A short oviduct extends from the ovary to the genital pore. A seminal receptacle may be present. The male genital pores may have several penile spines which may serve as a copulatory apparatus.

The embryology of the Kinorhyncha is unknown. Eggs hatch into minute juveniles which appear to have 11 segments. The juvenile passes through several successive stages by molting, then develops all 13 segments distinctly, and finally becomes an adult. Robert P. Higgins

Bibliography. B. Neuhaus and R. P. Higgins, Ultrastructure, biology, and phylogenetic relationships of Kinorhyncha, *Integr. Comp. Biol.*, 42:619–632, 2002.

Kirchhoff's laws of electric circuits

Fundamental laws dealing with currents at junctions and voltages around any loop within a network. These laws are commonly used in the analysis and solution of networks. They may be used directly to solve circuit problems, and they form the basis for network theorems used with more complex networks.

In the solution of circuit problems, it is necessary to identify the specific physical principles involved in the problem and, on the basis of them, to write equations expressing the relations among the unknowns. Physically, the analysis of networks is based on Ohm's law giving the branch equations, Kirchhoff's voltage law giving the loop voltage equations, and Kirchhoff's current law giving the node current equations. Mathematically, a network may be solved when it is possible to set up a number of independent equations equal to the number of unknowns. *See* CIRCUIT (ELECTRICITY); NETWORK THEORY.

When writing the independent equations, current directions and voltage polarities may be chosen arbitrarily. If the equations are written with due regard for these arbitrary choices, the algebraic signs of current and voltage will take care of themselves.

Kirchhoff's voltage law. One way of stating Kirchhoff's voltage law is: "At each instant of time, the algebraic sum of the voltage rise is equal to the algebraic sum of the voltage drops, both being taken in the same direction around the closed loop."

The application of this law may be illustrated with the circuit in **Fig. 1**. First, it is desirable to consider the significance of a voltage rise and a voltage drop, in relation to the current arrow. The following definitions are illustrated by Fig. 1.

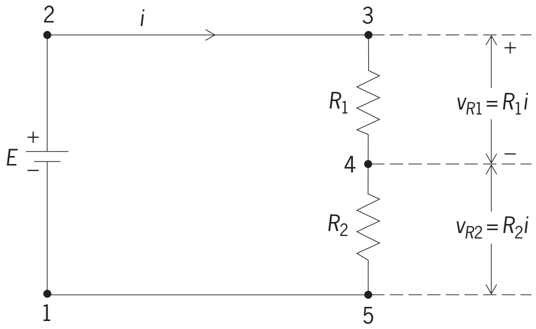


Fig. 1. Simple loop to show Kirchhoff's voltage law.

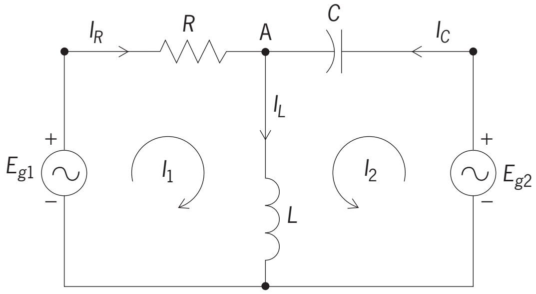


Fig. 2. Two-loop network demonstrating the application of Kirchhoff's voltage law.

A voltage rise is encountered if, in going from 1 to 2 in the direction of the current arrow, the polarity is from plus to minus. Thus, $v_{R1}; = R_1 i$ is a voltage drop from 3 to 4. The application of Kirchhoff's voltage law gives the loop voltage, Eq. (1).

$$E = v_{R1} + v_{R1} = R_1 i + R_2 i \tag{1}$$

In the network of **Fig. 2** the voltage sources have the same frequency. The positive senses for the branch currents I_R, I_L, and I_C are chosen arbitrarily, as are the loop currents I_1 and I_2. The voltage equations for loops 1 and 2 can be written using instantaneous branch currents, instantaneous loop currents, phasor branch currents, or phasor loop currents.

The loop voltage equations are obtained by applying Kirchhoff's voltage law to each loop as follows.

By using instantaneous branch currents, Eqs. (2) and (3) may be obtained. By using instantaneous loop

$$e_{g1} = Ri_R + L\frac{di_L}{dt} \tag{2}$$

$$e_{g2} = \frac{1}{C}\int i_C dt + L\frac{di_L}{dt} \tag{3}$$

currents, Eqs. (4) and (5) are obtained. Equations (6)

$$e_{g1} = Ri_1 + L\frac{d(i_1 + i_2)}{dt} \tag{4}$$

$$e_{g2} = \frac{1}{C}\int i_2\, dt + L\frac{d(i_2 + i_1)}{dt} \tag{5}$$

and (7) are obtained by using phasor branch

$$\mathbf{E}_{g1} = R\mathbf{I}_R + j\omega L\mathbf{I}_L \tag{6}$$

$$\mathbf{E}_{g2} = -j\frac{1}{\omega C}\mathbf{I}_C + j\omega L\mathbf{I}_L \tag{7}$$

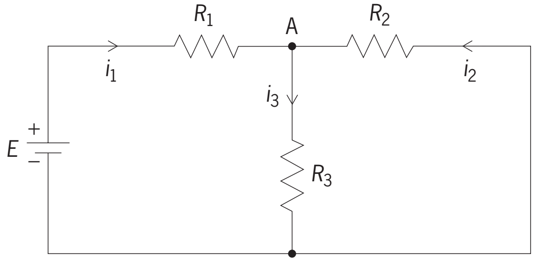


Fig. 3. Circuit demonstrating Kirchhoff's current law.

currents. By using phasor loop currents, Eqs. (8) and (9) may be obtained.

$$\mathbf{E}_{g1} = R\mathbf{I}_1 + j\omega L(\mathbf{I}_1 + \mathbf{I}_2) \tag{8}$$

$$\mathbf{E}_{g2} = -j\frac{1}{\omega C}\mathbf{I}_2 + j\omega L(\mathbf{I}_2 + \mathbf{I}_1) \tag{9}$$

See ALTERNATING-CURRENT CIRCUIT THEORY.

Kirchhoff's current law. Kirchhoff's current law may be expressed as follows: "At any given instant, the sum of the instantaneous values of all the currents flowing toward a point is equal to the sum of the instantaneous values of all the currents flowing away from the point."

The application of this law may be illustrated with the circuit in **Fig. 3**. At node *A* in the circuit in Fig. 3, the current is given by Eq. (10).

$$i_1 + i_2 = i_3 \tag{10}$$

The current equations at node *A* in Fig. 2 can be written by using instantaneous branch currents or phasor branch currents.

By using instantaneous branch currents, Eq. (11) is obtained.

$$i_R + i_C = i_L \tag{11}$$

By using phasor branch currents, Eq. (12) is obtained.

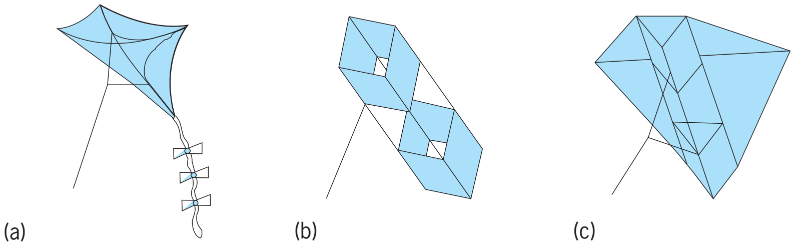
$$\mathbf{I}_R + \mathbf{I}_C = \mathbf{I}_L \tag{12}$$

See DIRECT-CURRENT CIRCUIT THEORY.

K. Y. Tang; Robert T. Weil

Kite

A tethered flying device that supports itself and the cable that connects it to the ground by means of the aerodynamic forces created by the relative motion of the wind. This relative wind may arise merely from the natural motions of the air or may be caused by towing the kite through the agency of its connecting cable.



Common forms of kites. (*a*) Bow kite. (*b*) Box kite. (*c*) Modified bow kite.

Kites take many forms (see **illus.**); the bow and box kites are common in the United States. In many countries, particularly in Asia, kites are frequently used in rituals and festivals; their bizarre forms and shapes are traditional, some having been developed centuries ago.

The lifting force of all kites is produced by deflecting the air downward, the resulting change in momentum producing an upward force. To be successful, a kite must have an extremely low wing loading (weight/area) so that it can fly even on days when the wind velocity is not high. It must be completely stable, since the only controls available to the operator are the length of cable and the rate at which it is taken in or let out. Efficient design requires that its lift-to-drag ratio be as high as possible. *See* AERODYNAMIC FORCE; AERODYNAMICS.

Experiments on the possible application of efficient aircraft-type lifting surfaces, in which most of the lift arises from the low pressures created by the air flowing over the upper surfaces, have shown them to be too sensitive to changes in wind force and direction. Under normal atmospheric conditions the use of this type of lifting surface results in a kite that behaves in a violent and unpredictable manner. For this reason, the higher drag associated with a surface from which the flow has separated is tolerated, and most of the lifting force is obtained from pressure on the lower surface, because stalled surfaces are much less sensitive to wind changes.

Both the lift-to-drag ratio and the stability of the kite are functions of the length of cable. The more cable released, the more drag created. The increased drag, combined with the increase in weight being supported, causes the kite to sag off downwind, reducing the flight angle, which is the angle formed between the horizontal and a line passing through the kite and the operator.

Most kites with a properly located cable pivot point, generally slightly ahead of the center of gravity, demonstrate longitudinal stability. Lateral and directional instabilities generally couple to produce violent motions. The longer the cable, the more these motions are damped. Lateral and directional stability are improved by the use of effective dihedral (the bow of the bow kite) and a flexible tail, which provides both directional stability and yaw damping.

David C. Hazen

Bibliography. R. Moulton, *Kites*, 1997; C. L. Strong, The lore and aerodynamics of making and flying kites, *Sci. Amer.*, 220(4):130–136, 1969; W. Yolen, *The Complete Book of Kites and Kite Flying*, 1976.

Kiwifruit

A vigorous deciduous fruiting vine (family Actinidiceae) that is native to central China, where it commonly grows in moist and sheltered areas on the forest edges. Kiwifruit requires both the female cultivar and a male pollenizer for successful fruit production. The kiwifruit industry depends on a single female cultivar, Hayward, the fruit having a creamy-white central core, black-brown seeds, and a bright translucent green outer flesh surrounded by a light brown fuzzy skin. It is adapted to moderate climates in the temperate zone and requires 600–850 h of winter chilling (temperatures between 32 and 45°F or 0 and 7°C) to ensure uniform budbreak. Kiwifruit wood is susceptible to winter injury at temperatures below 14°F (−10°C), and flower buds can be damaged by frost below 29°F (−1.5°C).

Origin and development. The kiwifruit (*Actinidia deliciosa*) was introduced into cultivation in New Zealand with seed brought from China in 1904, making it one of the most recently domesticated fruiting plants. It appears that all vines now in New Zealand descended from one male and two female plants. The principal kiwifruit-growing countries are Italy, New Zealand, Japan, Chile, France, Greece, United States (California), and Australia.

Kiwifruit is a source of vitamin C, minerals such as potassium, calcium, and phosphorus, and dietary fiber. The primary use is for the fresh market, although culled fruit is processed into canned and frozen fruit slices, wine, jam, juice, and dried products.

Cultivated kiwifruit vines have had very limited genetic selection and are very similar to those found in the wild. There are many species within the genus *Actinidia* with much variation (for example, fruit size and color) that is potentially useful for breeding programs and new introductions.

Propagation and cultivation. The Hayward cultivar is clonally propagated primarily by grafting onto seedling rootstock. Kiwifruit vines require deep, well-drained soils for best yields. Vines are grown on a trellis for support, and two shoots growing in opposite directions from a single straight trunk are trained to form a bilateral cordon structure. Spacing between rows and plants is usually about 15–16 ft (4.5–5 m). Adequate pollination is essential to produce large fruit, as the number of seeds directly correlates to fruit size. Vineyards are commonly planted at a ratio of one male plant to eight female plants. Beehives are used during bloom to increase pollination.

Fruit is borne on 1-year-old canes which arise from the supporting structure and are tied to the trellis wire. Female vines require extensive pruning to maintain strong annual cane growth, regulate crop yield and fruit size, and allow adequate light penetration for bud formation. Dormant pruning consists of removing fruiting canes back to 1-year-old replacement canes which will produce the following year's crop. Windbreaks are often used to shelter vines from the effects of wind damage: broken canes, lost crop, and scarred fruit. Vines have a high moisture requirement. When grown in areas with inadequate summer rainfall, they must have supplemental irrigation. Nitrogen, the most important nutrient, is applied annually during the growing season.

Harvesting and storage. Fruit is hand-picked in the fall once minimum maturity has been achieved. Packed fruit is stored at 32°F (0°C) for several months. During storage, the starch is converted to sugar and the fruit softens.

Diseases. Root and crown rot caused by several species of *Phytophthora* is a fungus disease that thrives in saturated soil conditions. Infected roots turn reddish brown, and the inner wood of infected crowns becomes dark brown. Vines may gradually decline or die within weeks of infection. Control methods include careful soil and water management to prevent the spread of disease from infested soil and irrigation or runoff water; and improved drainage.

Armillaria root rot is mainly a problem in New Zealand. Control methods, none of which are totally satisfactory, include fumigating infested soil, placing barriers in the soil between healthy and infected vines, and removing the remains of felled windbreak trees that harbor the disease.

Bacterial leafspot and blossom blight, caused by *Pseudomonas viridiflava*, is spread to floral buds and young leaves during spring rains. Infected buds may turn brown, fail to open or develop fully, and ultimately drop.

Storage rot caused by *Botrytis cinerea* is the most serious of postharvest diseases. The fungus colonizes stem scars or any damaged area usually after several weeks of storage and may spread to adjacent healthy fruits. Control is based on using preharvest fungicide sprays, avoiding soft fruit through proper cultural practices, and proper handling during harvest and storage. *See* FRUIT; PLANT PATHOLOGY; THEALES.

Janine K. Hasey

Bibliography. K. E. Conn, W. D. Gubler, and J. K. Hasey, Pathogenicity and relative virulence of nine *Phytophthora* spp. from kiwifruit, *Phytopathology*, 81:974–979, 1991; J. K. Hasey and R. S. Johnson (eds.), *Kiwifruit Growing and Handling*, 1994; I. J. Warrington and G. C. Weston (eds.), *Kiwifruit Science and Management*, 1990.

Klebsiella

A genus of gram-negative, nonmotile bacteria whose members ferment lactose, do not decarboxylate ornithine or form hydrogen sulfide, and often produce large mucoid colonies and gas from glucose. Characteristic large mucoid colonies are due to production of a large amount of capsular material. Some isolates, particularly those from urinary tract infections, produce a slow-acting urease enzyme. Species of *Klebsiella* are commonly found in soil and water, on plants, and in animals and humans. Harmless strains of *Klebsiella* are beneficial because they fix nitrogen in soil. Pathogenic species include *K. pneumoniae, K. rhinoscleromatis*, and *K. ozaenae*, also known as *K. pneumoniae* subspecies *pneumoniae, rhinoscleromatis*, and *ozaenae*. Species and strains can be differentiated by biochemical reactions and typed by bacteriocin susceptibility and serologically by antigen.

Klebsiella pneumoniae is the second most frequently isolated colon-related bacterium in clinical laboratories. The carbohydrate-containing capsule of *Klebsiella* promotes virulence by protecting the encased bacteria from ingestion by leukocytes; nonencapsulated variants of *Klebsiella* do not cause disease. Capsular types 1 and 2 cause pneumonia, accounting for about 3% of all life-threatening acute bacterial pneumonias. Capsular types 8, 9, 10, and 24 are commonly associated with urinary tract infections, second in prevalence only to *Escherichia coli*. *See* ESCHERICHIA; PNEUMONIA.

Klebsiella accounts for a large percentage of hospital-acquired infections, mostly skin infections (in immunocompromised burn patients), bacteremia, and urinary tract infections. It is also the most common contaminant of intravenous fluids such as glucose solutions and other medical devices. Partly, this problem is due to the fact that from 5 to 10% of healthy individuals carry *Klebsiella* in their respiratory tract. *See* HOSPITAL INFECTIONS.

Klebsiella may produce *E. coli*–like enterotoxins and cause acute gastroenteritis in infants and young children. Enteric illnesses due to *Klebsiella* are more predominant where populations are more crowded and conditions less sanitary. Other virulence factors of *Klebsiella* include a relatively high ability to survive and multiply outside the host in a variety of environments, and its relatively simple growth requirements. An important virulence factor of *Klebsiella* is its pyrogenic, or fever-producing, endotoxin, which is a heat-stable lipopolysaccharide component of the organism's cell wall. *Klebsiella* also has a relatively nonspecific virulence factor, which is the so-called common pili produced by almost all enteric bacteria. Common pili are nonflagellar hairlike appendages which function as receptors for mannose, a carbohydrate which is almost universally present on the surface of animal cells. These appendages function as adhesins and are thus important in the colonization phase of infection and in maintaining and spreading infection. *See* ENDOTOXIN.

Klebsiella rhinoscleromatis causes rhinoscleroma, a chronic destructive granulomatous disease of the upper respiratory tract that is most common in eastern Europe, central Africa, and tropical South America. *Klebsiella ozaenae* is one cause of chronic rhinitis (ozena), a destructive atrophy of the nasal mucosa, and is infrequently isolated from urinary tract infections and bacteremia. *See* MEDICAL BACTERIOLOGY.

Doyle J. Evans, Jr.

Bibliography. S. Baron (ed.), *Medical Microbiology*, 4th ed., 1996.

Klystron

An evacuated electron-beam tube in which an initial velocity modulation imparted to electrons in the beam results subsequently in density modulation of the beam. A klystron is used either as an amplifier in the microwave region or as an oscillator. For use as an amplifier, a klystron receives microwave energy at an input cavity through which the electron beam passes. The microwave energy modulates the velocities of electrons in the beam, which then enters a drift space. Here the faster electrons overtake the slower to form bunches. In this manner, the uniform current density of the initial beam is converted to an alternating current. The bunched beam with its significant component of alternating current then passes through an output cavity to which the beam transfers its ac energy. *See* MICROWAVE.

Klystron amplifier. In a typical klystron (**Fig. 1**), a stream of electrons from a concave thermionic cathode is focused into a smaller cylindrical beam by the converging electrostatic fields between the anode, cathode, and focusing electrode. The beam passes through a hole in the anode and enters a magnetic field parallel to the beam axis. The magnetic field holds the beam together, overcoming the electrostatic repulsion between electrons which would otherwise make the beam spread out rapidly. The electron beam goes through the cavities of the klystron in sequence, emerges from the magnetic field, spreads out, and is stopped in a hollow collector where the remaining kinetic energy of the electrons is dissipated as heat. *See* ELECTRON MOTION IN VACUUM.

The signal wave to be amplified is introduced into the first, or buncher, cavity through a coaxial transmission line. This hollow metal cavity is a resonant circuit, analogous to the familiar inductance-capacitance combination, with the electric field largely concentrated in the reentrant noses so that the highest voltage occurs between them. The inductance may be considered as a single-turn conductor formed by the outer metal walls. In Fig. 1 the current in the center conductor of the input transmission line flows through a loop inside the cavity and back to the outer conductor. The magnetic flux generated in the loop links through the cavity inductance, as in a transformer. At the resonant frequency of the cavity, the voltage across the reentrant section through which the electron beam passes is built up by the cavity configuration to 10–100 times the voltage in the input line. *See* CAVITY RESONATOR.

Figure 2 shows the pattern of the electric field in the cavity and how it varies cyclically with time. As electrons pass through the gap in which the cavity field is concentrated (Fig. 2*a*), they are accelerated or decelerated (Fig. 2*b*), depending on the instantaneous direction of the field.

Figure 3 illustrates graphically the effect of these velocity changes. Each slanted line represents the flight of an electron as a function of time measured in electrical degrees as the electron travels from the buncher gap. The slope of a line is thus the velocity of that electron. The velocities leaving the buncher vary sinusoidally with time, as determined by the instantaneous field. The horizontal broken line represents a fixed point beyond the buncher. The flow

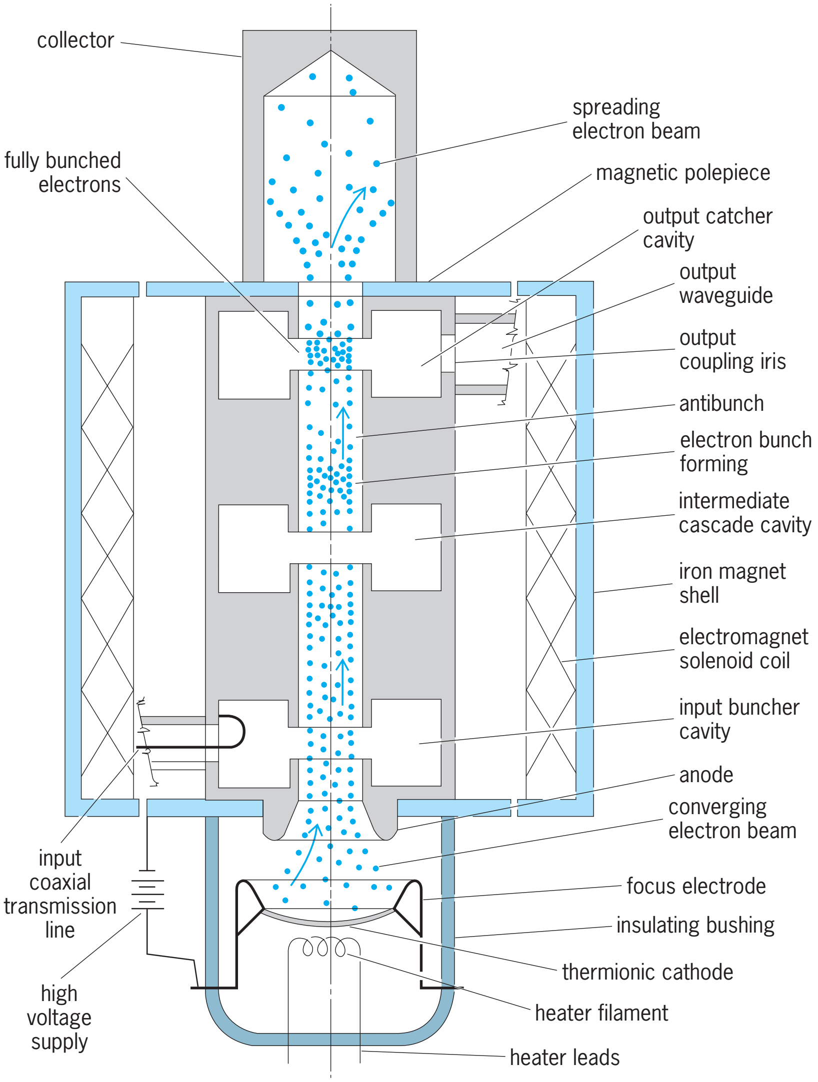


Fig. 1. Cross section of cascade klystron amplifier. (*Varian Associates*)

of electrons past this point is given by the time sequence in which the electron paths cross the broken line. Figure 3 shows how the electrons have gathered into bunches. The rate of current flow is now periodic with time so that the current has an alternating component.

When the bunched beam passes through a second cavity, its space charge induces in the walls of this cavity an alternating current of opposite sign to the electron current. The cavity is tuned to the input frequency so that it has a high resonant impedance; the induced current flowing through this impedance generates voltage in the cavity.

In Fig. 1 the second cavity is not coupled to any outside circuits. Voltage built up here by the beam current produces further velocity modulation in the beam. The resulting alternating current component is about 10 times greater than the initial current. More of these uncoupled cascade cavities can be added for increased amplification.

The final output cavity is coupled into a transmission line (a waveguide in Fig. 1) which carries off the generated power to its useful destination. Because the cavity is tuned to resonance, its reactance is canceled. The induced current flowing through its pure resistive impedance generates in-phase voltage in the direction opposing the current flow. Thus the field in the gap is at its maximum decelerating value at the time a bunch of electrons passes. Most of the electrons therefore are slowed down, and there is a net transfer of kinetic energy of the electrons into electromagnetic energy in the cavity. Klystron amplifiers are used in transmitters for radar and one-way radio communication, for driving particle accelerators, and for dielectric heating. The useful range of frequencies is from 400 MHz to 40 GHz. Power levels range from a few watts up to 400 kW of continuous power or 20 MW for short pulses. Amplification is about 10 dB for a two-cavity tube. With more cavities, gains up to 60 dB are practical.

Figure 4 shows the construction of a four-cavity amplifier rated at 2 MW pulsed output at 2.8 GHz. It operates in a solenoid magnet, as in Fig. 1. The

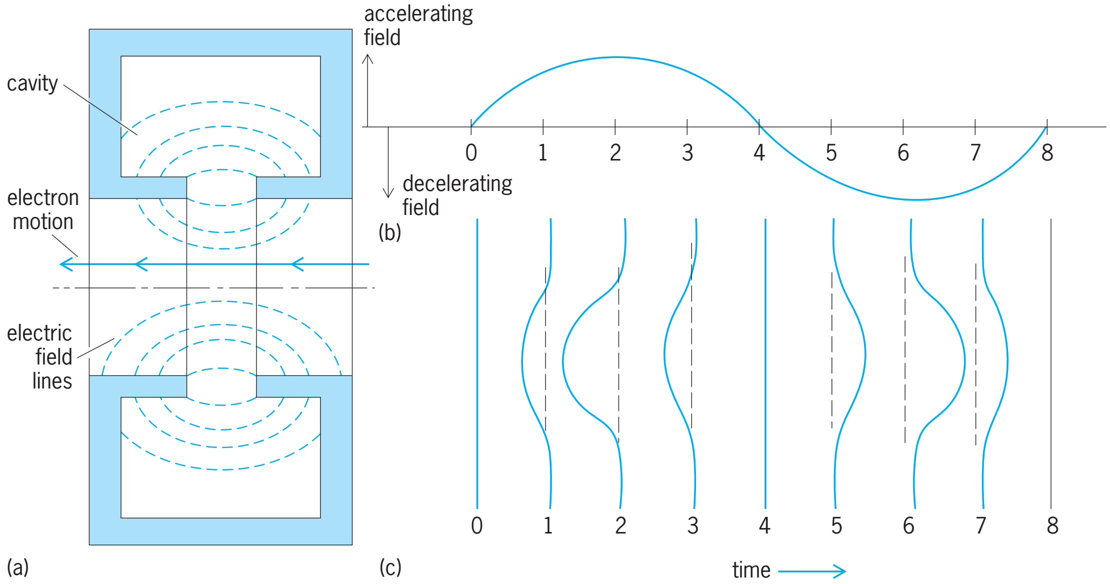


Fig. 2. Cavity concentrates electric field between the reentrant noses. As field varies sinusoidally with time, an electron crossing in the gap between the noses experiences an electric field whose strength and direction depend on the instantaneous phase of the field. (*a*) Map of instantaneous field. (*b*) Cycle variation of field. (*c*) Profile of field on the axis at various times in the cycle. (*Varian Associates*)

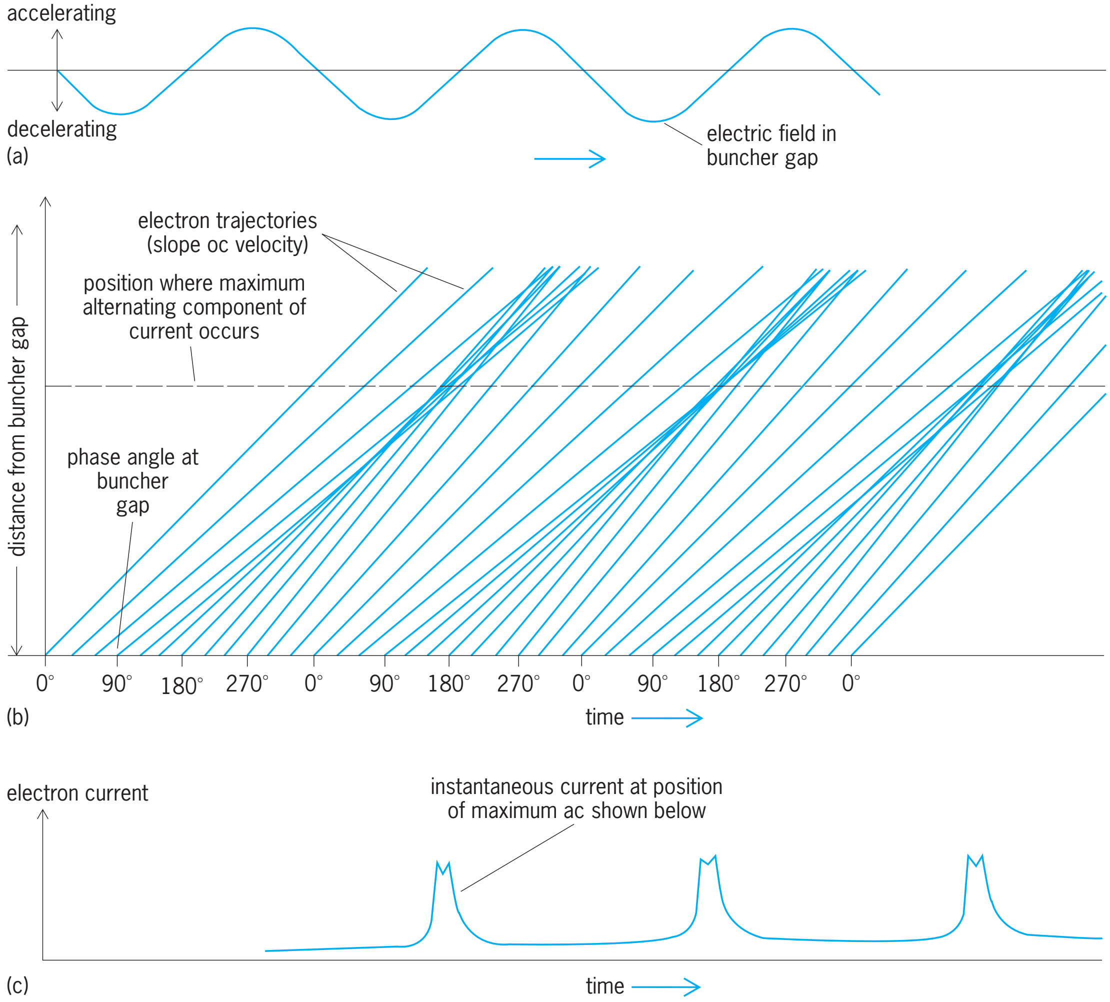


Fig. 3. Bunching the electron stream. (*a*) As the electric field varies periodically, electrons traversing the cavity are speeded up or slowed down. (*b*) Distance-versus-time lines graph the gradual formation of bunches. (*c*) Current passing a fixed point becomes periodic, that is, it becomes alternating. (*Varian Associates*)

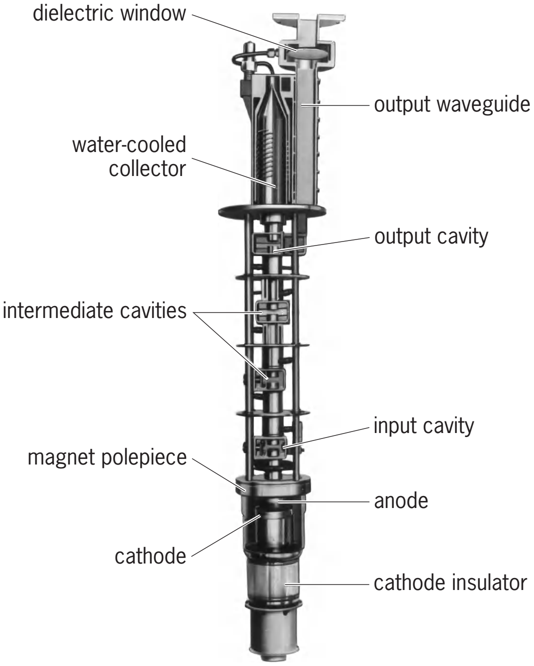


Fig. 4. Cutaway view of four-cavity amplifier. (*Varian Associates*)

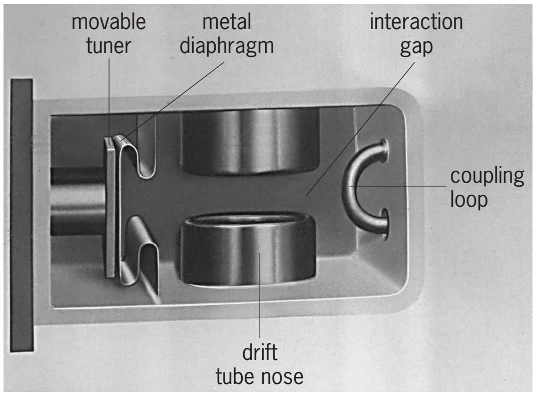


Fig. 5. Detail of input cavity for four-cavity klystron amplifier of Fig. 4. (*Varian Associates*)

cavities are tuned to the operating frequency by moving one flexible inner wall of the box-shaped cavity, changing its volume and its effective inductance. In **Fig. 5** the details of the input cavity are enlarged.

Reflex oscillator. Klystrons may be operated as oscillators by feeding some of the output back into the input circuit. More widely used is the reflex oscillator in which the electron beam itself provides the feedback. **Figure 6** illustrates the operation. The beam is focused through a cavity, as in the amplifier. No magnetic field is needed to keep the beam focused because the total travel distance is short and the amount of natural spreading is tolerable. The cavity usually has grids with open mesh through which the electrons can penetrate. The purpose of the grids is to concentrate the electric field in a short space so that the field can interact with a slow, low-voltage electron beam.

In the cavity the beam is velocity-modulated as in the amplifier. Leaving the cavity, the beam enters a region of dc electric field opposing its motion, produced by a reflector electrode operating at a potential negative with respect to the cathode. The electrons do not have enough energy to reach the electrode, but are reflected in space and return to pass through the cavity again. The points of reflection are determined by electron velocities, the faster electrons going farther against the field and hence taking longer to get back than the slower ones.

A trajectory plot for the reflex oscillator is shown in **Fig. 7**. In a uniform retarding field the space-versus-time curves are parabolas. As in Fig. 3, velocity modulation produces bunches of electrons. If the voltages are adjusted so that the average time to return is $n + {}^{3}\!/_{4}$ cycles (n = integer), the bunches cross the cavity when the alternating field is maximum in the decelerating direction. This transfers beam energy to the cavity.

Because the reflex klystron has only one cavity, it is easy to tune its frequency with a single adjustment. Power output is from 10 mW to a few watts. Reflex oscillators are used as signal sources from 3 to 200 GHz, as the transmitter tubes in line-of-sight radio relay systems, and in low-power radars.

A cutaway of a typical reflex klystron is shown in **Fig. 8**. This tube is tuned by deforming the upper

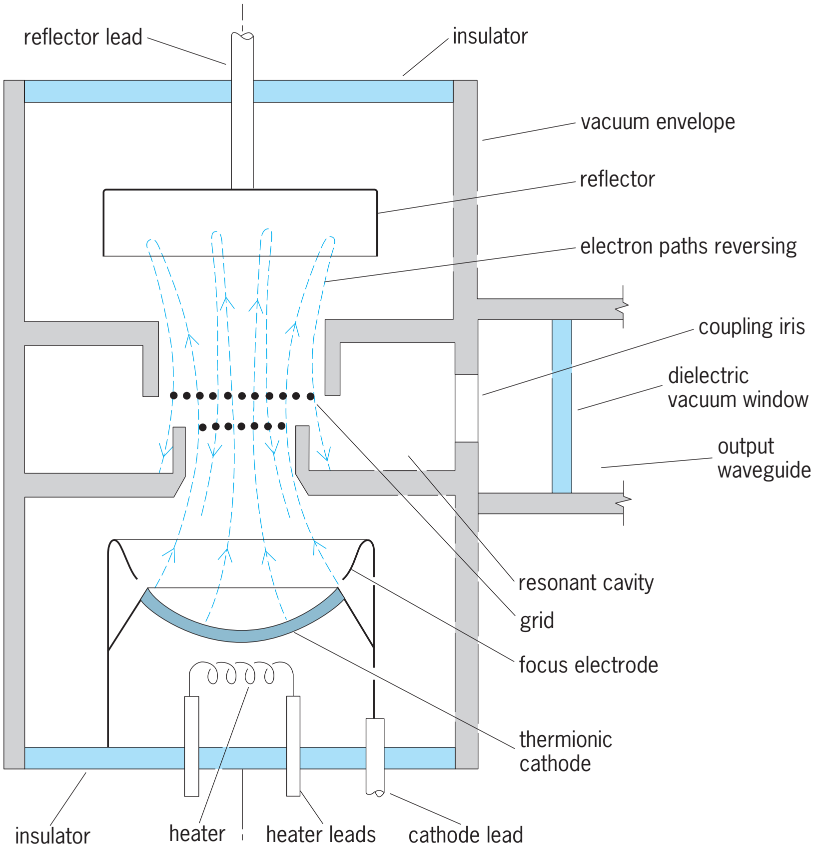


Fig. 6. Schematic cross section of reflex oscillator. (*Varian Associates*)

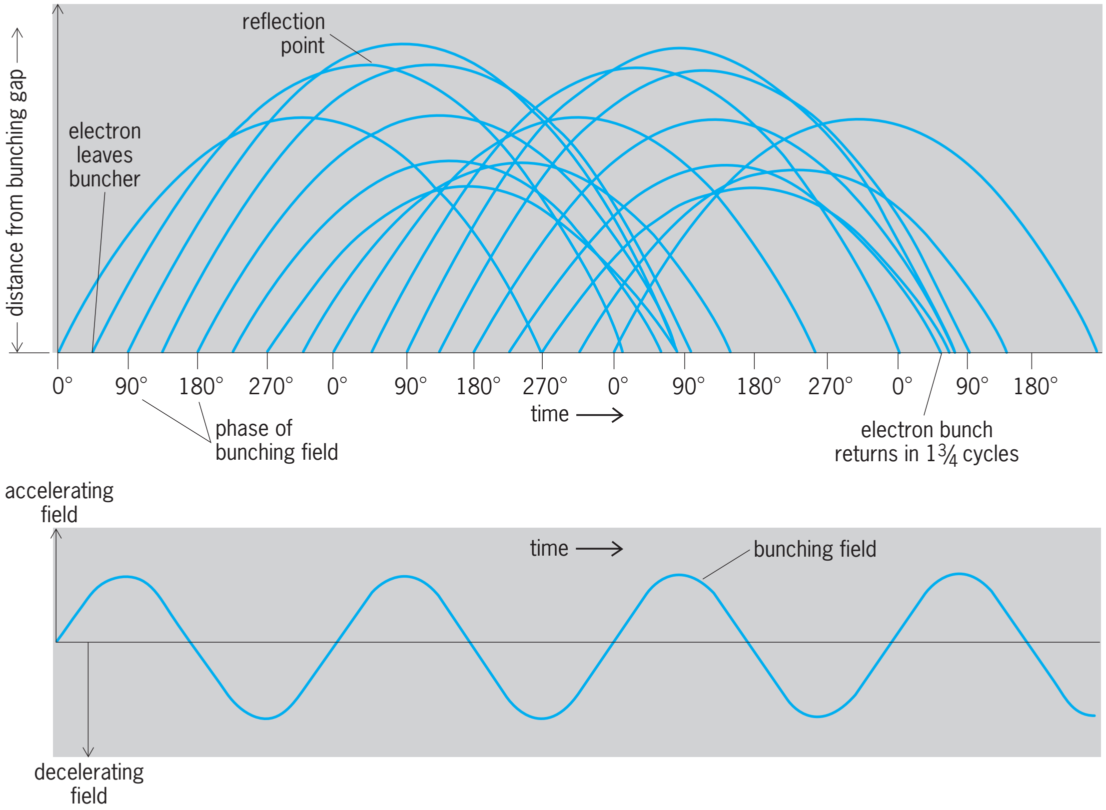


Fig. 7. Trajectories show bunching in a reflex oscillator. Electrons are turned back by a retarding field, faster ones going farther and taking longer. The bunched beam returns through the cavity. (*Varian Associates*)

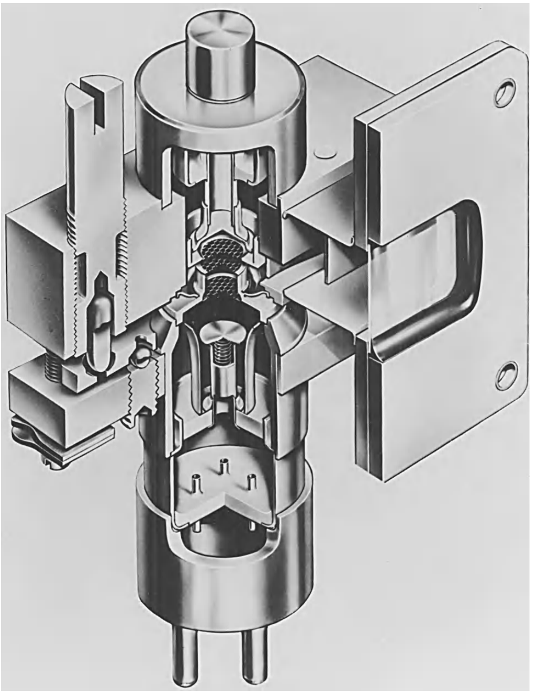


Fig. 8. Cutaway of reflex oscillator shows (at left) the screw that deforms the flexible bottom wall of the cavity to change its resonant frequency. The beam passes through honeycomb grids in the cavity. The cathode leads come out at the bottom socket, and the reflector lead comes out at the top. The output waveguide has a mica vacuum window. (*Varian Associates*)

cavity wall, which varies the spacing between the grids and hence the effective capacitance of the cavity resonator. Richard B. Nelson

Bibliography. J. J. Carr, *Elements of Microwave Technology*, 1989; K. Chang, *Handbook of Microwave and Optical Components*, 4 vols., 1990; A. S. Gilmour, *Microwave Tubes*, 1986; T. K. Ishii, *Practical Microwave Electron Devices*, 1990; S. Y. Liao, *Microwave Devices and Circuits*, 3d ed., 1990; D. M. Pozar, *Microwave Engineering*, 2d ed., 1997.

Knudsen number

In gas dynamics, the ratio of the molecular mean free path λ to some characteristic length L: $\mathrm{Kn} = \lambda/L$. The length chosen will depend on the problem under consideration. It may be, for example, the diameter of a pipe or an object immersed in a flow, or the thickness of a boundary layer or a shock wave. *See* MEAN FREE PATH.

The magnitude of the Knudsen number determines the appropriate gas dynamic regime. When the Knudsen number is small compared to unity, of the order of $\mathrm{Kn} \leq 0.1$, the fluid can be treated as a continuous medium and described in terms of the macroscopic variables: velocity, density, pressure, and temperature. In the transition flow regime, for Knudsen numbers of the order of unity or greater, a microscopic approach is required, wherein the trajectories of individual representative molecules are considered, and macroscopic variables are obtained from the statistical properties of their motions. In both internal and external flows, for Kn ≥ 10, intermolecular collisions in the region of interest are much less frequent than molecular interactions with solid boundaries, and can be ignored. Flows under such conditions are termed collisionless or free molecular. In the range $0.1 \leq \text{Kn} \leq 1.0$, termed the slip flow regime, it is sometimes possible to obtain useful results by treating the gas as a continuum, but allowing for discontinuities in velocity and temperature at solid boundaries.

Since the mean free path is inversely proportional to the gas density, departures from continuum gas dynamics are usually associated with flows in vacuum systems and in high-altitude aerodynamics, often referred to as rarefied gas dynamics. However, such departures can also occur at normal densities owing to the existence of a very small characteristic length, as in the motion of very small particles such as aerosols, and in the flow through very narrow passageways. With the ever-increasing power of computational resources, numerical methods have displaced analytical methods as the predominant approach to the treatment of flows with complicated geometries and complex chemistry, over the entire range of Knudsen number. Monte Carlo methods are particularly useful in the transition flow regime. *See* COMPUTATIONAL FLUID DYNAMICS; GAS DYNAMICS; KINETIC THEORY OF MATTER; MONTE CARLO METHOD; RAREFIED GAS FLOW. Lawrence Talbot

Bibliography. G. A. Bird, *Molecular Gas Dynamics and the Direct Simulation of Gas Flows*, Oxford University Press, 1994.

Koala

A single species, *Phascolarctos cinereus*, which is a member of the family Phalangeridae in the mammalian order Marsupialia (pouch-bearing animals). It is a small, clean, affectionate animal that weighs from 11 to 17 lb (5 to 8 kg) when mature. Koalas are restricted to eastern Australia, where the eucalyptus grows. Not only do they have a specialized diet of eucalyptus leaves, but the leaves must be of a certain age from a specific species of tree, and the tree must grow upon a certain type of soil. Each adult will eat an average of 2–3 lb (0.9–1.5 kg) of leaves each day. While the diet is a definite limiting factor to an increase in the population, millions died from epidemic diseases at the turn of the twentieth century, and later thousands were killed for their fur. These animals are now strictly protected by the Australian government because their numbers are small. Exportation is prohibited.

The koala is well adapted to its arboreal habitat, since both the fore- and hindlimbs have two of the clawed digits opposing the other three, thus constituting an efficient grasping organ (see **illustration**). The koala breeds once each season, and the usual number of offspring is one. It remains in the pouch for 6 months; then it clings to the back of its mother and is carried around in this manner until 1 year old. *See* MAMMALIA; MARSUPIALIA. Charles B. Curtin

Koala (*Phascolarctos cinereus*), a slow-moving arboreal animal. (*Photo by Gerald and Buff Corsi; © 2002 California Academy of Sciences*)

Bibliography. R. M. Nowak, *Walker's Mammals of the World*, Johns Hopkins University Press, 1999.

Kohlrabi

A cool-season biennial crucifer, *Brassica caulorapa* and *B. oleracea* var. *caulo-rapa*, of northern European origin belonging to the plant order Capparales. Kohlrabi is grown for its turniplike enlarged

Kohlrabi (*Brassica caulorapa*), cultivar Early White Vienna. (*Joseph Harris Co., Rochester, New York*)

stem, which is usually eaten as a cooked vegetable (see **illus.**). Kohlrabi is a German word meaning cabbage-turnip and reflects a similarity in taste and appearance to both vegetables. Cultural practices for kohlrabi are similar to those used for turnips. White Vienna and Purple Vienna are popular varieties (cultivars). Harvesting, when the enlarged stems are 2–3 in. (5–8 cm) in diameter, is usually 2 months after planting. A common cooked vegetable in Europe, especially Germany, kohlrabi is of minor importance in the United States. *See* CABBAGE; CAPPARALES; KALE; TURNIP.

H. John Carew

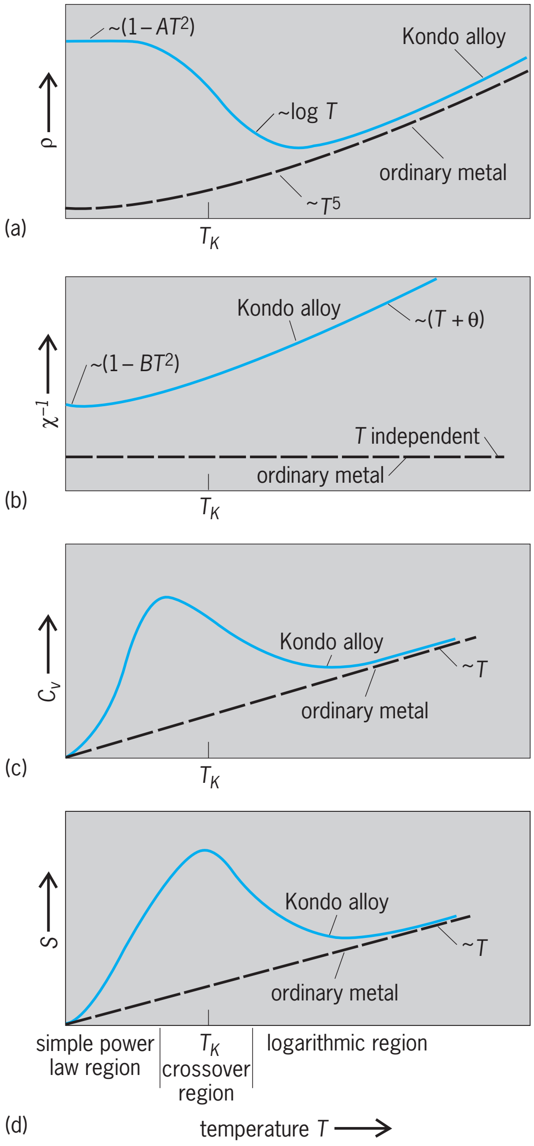
Kondo effect

An unusual, temperature-dependent effect displayed in the thermal, electrical, and magnetic properties of nonmagnetic metals containing very small quantities of magnetic impurities. A striking example is the anomalous, logarithmic increase in the electrical resistivity ρ with decreasing temperature. Other properties, such as heat capacity C_v, magnetic susceptibility χ, and thermoelectric power S, also display anomalous behavior because of the Kondo effect. For these properties, the temperature dependence of a typical dilute magnetic metal (Kondo alloy) differs greatly from the behavior expected of an ordinary metal containing no magnetic impurities (see **illus.**).

The Kondo effect has been observed in a wide variety of dilute magnetic alloys. Usually these alloys are made from a nonmagnetic host such as copper, silver, gold, magnesium, or zinc and a small amount of a magnetic metal impurity such as chromium, manganese, iron, cobalt, nickel, vanadium, or titanium. Typical concentrations range from about one to a few hundred magnetic atoms per million host atoms. At higher concentrations, the dilute magnetic alloys may display spin-glass behavior. *See* SPIN GLASS.

Importance. The Kondo effect is used in thermometry applications, especially thermocouple thermometers at very low temperatures (that is, millikelvin temperatures). In other applications where the properties of pure metals are studied, the Kondo effect serves as a useful indicator of the metal's magnetic-impurity level.

The problem of understanding the Kondo effect is considered important since it is recognized to be a simpler version of the more complex problem of understanding ferromagnetism in magnetic materials, which is one of the great challenges in physics. Basically the Kondo effect is an example of the most simple possible magnetic system—a single magnetic atom in a nonmagnetic environment. (The alloys used are so dilute that the interaction between different magnetic impurities can be safely ignored.) Although this involves a simple physical model, the problem has required some of the most sophisticated mathematical techniques known to advance its understanding.



Temperature dependence of a typical Kondo alloy compared with an ordinary metal for four experimentally measurable quantities: (*a*) electrical resistivity ρ, (*b*) inverse magnetic susceptibility χ^{-1}, (*c*) thermodynamic heat capacity C_v, and (*d*) thermoelectric power *S*. Characteristic dependences on temperature *T*, the Kondo temperature T_K, and three regions of behavior are identified.

An important step in this direction was the development of a partial mathematical solution of the Kondo problem using renormalization field theory techniques. Information gained in this step helped with the final development of a mathematically exact solution of the Kondo problem. The exact solution permits a systematic calculation of all properties (resistivity, thermal conductivity, thermopower, specific heat, magnetic susceptibility, neutron scattering behavior, and so forth) and provides a physical understanding of these properties. The theoretical work on the Kondo problem has been connected with new understanding in a variety of other scientific disciplines such as condensed-matter physics, surface physics, critical phenomena, elementary particle physics, magnetism, molecular physics, and chemistry, where parallels and analogs to the Kondo problem can be identified and utilized. *See* CRITICAL PHENOMENA; FERROMAGNETISM; RENORMALIZATION.

Effect of exchange interaction. The first atomic understanding of the phenomenon was achieved in 1964, when Jun Kondo attempted to explain the puzzling minimum in the resistivity of dilute magnetic alloys at low temperatures. Kondo analyzed the simple model of a single magnetic impurity in a nonmagnetic metal by assuming an exchange interaction between the impurity atom and conduction electrons. The exchange interaction J, as it is known in quantum mechanics, is a fundamental consequence of basic symmetry relations that all atomic particles must obey, as a result of their intrinsic spin. By calculating how much the mobile conduction electrons were scattered off the fixed magnetic impurities, Kondo was able to determine the electrical resistivity. He found that the scattering cross section of the electrons showed a remarkable increase (mathematically a logarithmic singularity) at low temperatures when the analysis was carried one step beyond the conventional quantum-mechanical calculation. *See* EXCHANGE INTERACTION.

At a certain characteristic temperature, now referred to as the Kondo temperature T_K, the Kondo effect becomes clearly observable. In Kondo's theory this temperature was identified when the mathematical series of terms in the quantum-mechanical calculation diverged logarithmically. Later work showed that T_K depended on the exchange interaction J, the number of quantum states per unit energy n (often called density of states), and the kinetic energy of the free conduction electrons ϵ_F (more explicitly, the Fermi energy). More refined theories show that a suitable mathematical expression for T_K is

$$T_K \approx \epsilon_F J^{1/2} \exp\left(-\frac{1}{nJ}\right)$$

See FERMI SURFACE; FREE-ELECTRON THEORY OF METALS.

For a given alloy, the values of J, n, and ϵ_F are unique and cannot be adjusted; therefore, at temperatures much higher than T_K the metal atoms experience random thermal agitations that overwhelm the exchange interaction and the Kondo effect becomes insignificant. As temperatures approach T_K, just the reverse occurs. The thermal agitation drops dramatically with falling temperature, and the exchange interaction, which is temperature-independent, dominates; hence the Kondo effect becomes observable (see illus.).

In early experimental investigations (beginning in the 1930s), the Kondo effect was more easily detected in alloys at rather low temperatures (a few hundredths to a few degrees above absolute zero) where the thermal agitation was assuredly small. However, with improved measuring techniques and better alloys, the Kondo effect has been observed at much higher temperatures, corresponding to T_K of several hundred kelvins, well above room temperatures.

Many-body effects. Although the logarithmic rise in the resistivity could be explained, there was still the problem of finding a theory that accounted for the leveling-off of the resistivity at the lowest temperatures (see illus.). This problem was resolved by realizing that a type of many-body effect takes place between the conduction electrons and a magnetic impurity atom, leading to an effective coupling between the impurity and its neighboring electrons. This many-body effect concerns the problem of electrons in a metal experiencing a complex environment where many particles may interact simultaneously around an impurity. In a simple way, the effective coupling may be viewed as a cloud of interactions formed by compounding many simple two-body exchange interactions. At higher temperatures the effective coupling is weak because the thermal agitation rapidly destroys the simple exchange interaction between an electron and a magnetic impurity before another electron, in the vicinity of the impurity, has a chance to react to the effect of the interaction. At lower temperatures the effective coupling becomes more prominent since the interactions are not as rapidly destroyed by thermal agitation. (Technically the increase in coupling is a consequence of an increasing correlation between the spin polarization of the conduction electrons and the impurity spins.) Eventually the coupling gains sufficient strength that the impurity's magnetic spin becomes screened by the magnetic spin of the electron. At this point, there no longer appear to be localized centers of magnetic atoms, and the electron-scattering cross section becomes constant; consequently, the logarithmic rise in the resistivity ceases. *See* SPIN (QUANTUM MECHANICS).

Crossover problem. Bridging the crossover from weak coupling to strong coupling presented a major challenge to many-body physics, and this difficult task was first accomplished in 1973 by Kenneth Wilson through computer calculations based on renormalization group theory. In 1981 this crossover problem was solved by a different method, which yielded an analytical solution. The agreement between these two approaches was considered excellent. There also was the advantage that the analytical method provided a calculated expression for the magnetic susceptibility as observed in the Kondo effect. However, these approaches were still limited by their inability to calculate the temperature dependence of the various properties associated with the Kondo effect.

Exact solution. An important breakthrough in the problem came with the study of dilute impurity cases which have high magnetic impurity concentrations. Especially interesting were alloys containing rare-earth elements such as dilute cerium in a metallic host of lanthanum hexaboron (LaB_6). Such materials are generally classed as heavy fermions or mixed-valence systems. In seeking an explanation of these systems, there arose the realization that the large degeneracy number N of an impurity atom's states (for example, the states of the *f*-orbital electrons of cerium with the same energy) could be used to develop mathematical expressions that describe appropriate physical properties. In particular, these mathematical expressions were found to be summations of a convergent series of terms in inverse powers of N, leading subsequently to a formalism known as the $1/N$ expansion method. Coupled with other mathematical formalisms, the $1/N$ expansion method, as applied to the Kondo problem, has been used to develop a consistent set of integral equations that can be solved numerically. Specifically, the computation for one alloy (dilute cerium in lantharum hexaboron) has been carried to completion. In terms of only a single parameter (the peak position of the Kondo resonance T_0 which is proportional to the Kondo temperature T_K), the theory explains the Kondo behavior of this alloy and fits the data of four different Kondo-effect properties, each measured over four decades of temperature. This achievement is considered to be an exact solution of the Kondo problem, and one that represents a universal understanding of various physical properties displaying the Kondo effect.

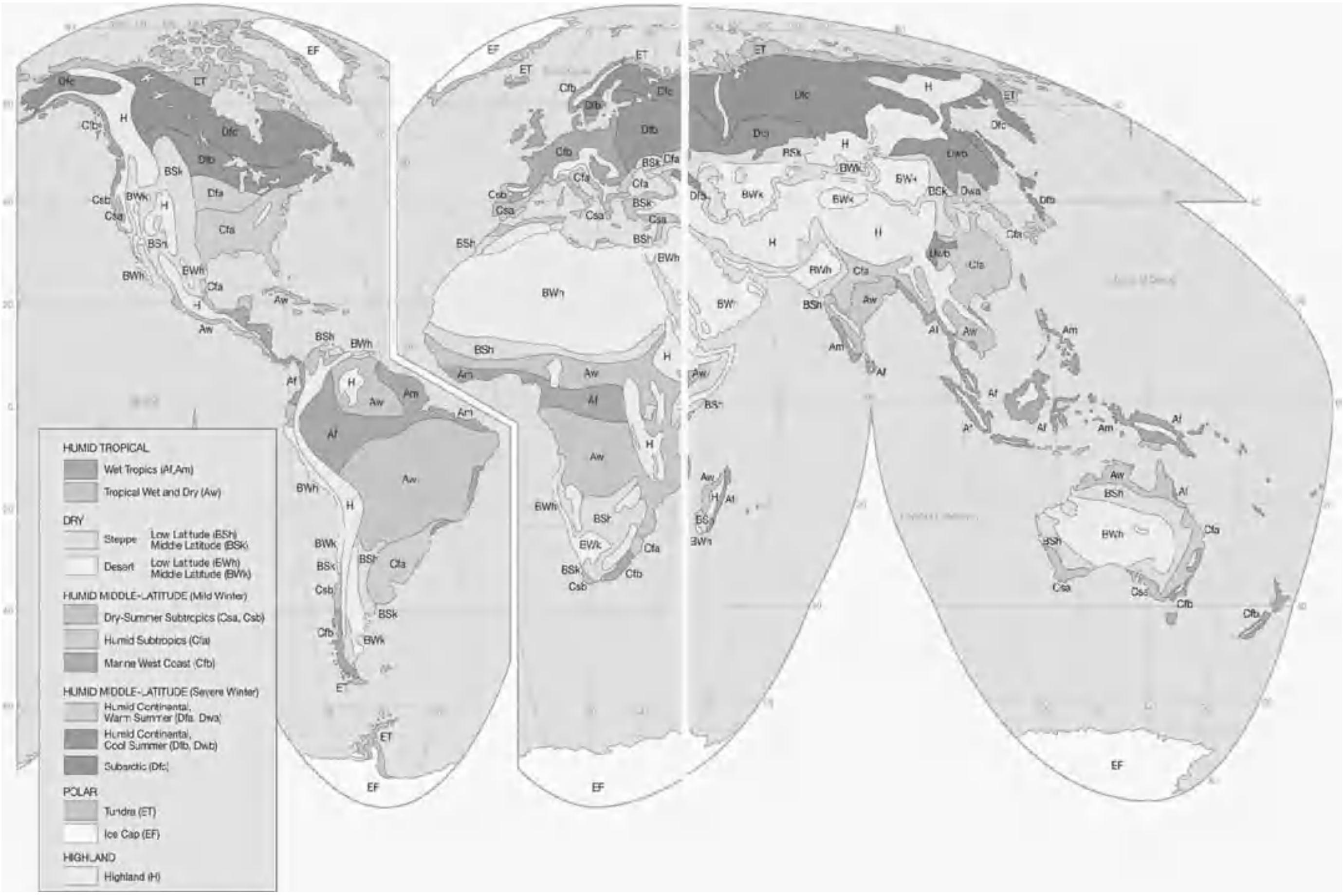
Many generalizations of the Kondo problem and its relatives, that is, specialized impurity model problems, continue to be a focus of interest. In one special case, namely, the general Kondo problem, success in finding an exact solution has been achieved by using a conformal field theory approach. In this approach the impurity is identified as an immobile quasiparticle while the scattering electron is equated with a mathematical artifice known as a kink, which is a special type of field configuration interpolating between adjacent potential energy wells. *See* ELECTRICAL RESISTIVITY; HEAT CAPACITY; MAGNETIC SUSCEPTIBILITY; SPECIFIC HEAT OF SOLIDS; THERMOELECTRICITY.

Wiley P. Kirk

Bibliography. N. Andrei, K. Furuya, and J. H. Lowenstein, Solution of the Kondo problem, *Rev. Mod. Phys.*, 55:331–402, April 1983; P. Fendley, Kinks in the Kondo problem, *Phys. Rev. Lett.*, 71:2485–2488, 1993; A. C. Hewson, *The Kondo Problem to Heavy Fermions*, 1997; R. M. White, *Quantum Theory of Magnetism*, 1983; J. W. Wilkins, Solution of the Kondo problem (at last), *Phys. Today*, 39(1):22, January 1986; K. G. Wilson, The renormalization group: Critical phenomena and the Kondo problem, *Rev. Mod. Phys.*, 47:773–840, October 1975.

Köppen climate classification system

A system for assigning short codes to the Earth's climate zones. It consists of a hierarchical arrangement of the major climatic groups, subcategories within



Generalized Köppen climate classification map. (*From E. Tarbuck and F. K. Lutgens, Earth Science, 11th ed., Prentice Hall, 2006*)

Köppen system

First division	Description	Second division	Description	Third division	Description
A	**Tropical** Coolest month's average temperature >64°F/180°C	m	A period of less precipitation in an otherwise very moist climate		
		w	Dry winters, wet summers		
		f	Constantly wet (no less than 2.4 in./60 mm of rain, in the driest month)		
B	**Dry** Potential evaporation and transpiration exceed precipitation	S	More precipitation than deserts, but under 20 in./500 mm per year (semiarid/steppe)	h	Mean annual temperature >64°F/18°C (tropical)
				k	Mean annual temperature <64°F/18°C (temperate)
		W	Less than 10 in./250 mm of precipitation per year (arid/desert)	h	Mean annual temperature >64°F/18°C (tropical)
				k	Mean annual temperature <64°F/18°C (temperate)
C	Midlatitude with mild winters	s	Mild with dry summer (mediterranean)	a	Hot summer season
				b	Warm summer season
		f	Mild with no dry season (humid subtropical)	a	Hot summer season
				b	Warm summer season
				c	Cool summer season
		w	Mild with dry winter	a	Hot summer season
D	Midlatitude with severe winters	f	Humid with severe winters; no dry season (humid-continental)	a	Hot summer season
				b	Warm summer season
				c	Cool summer season
				d	Cool summers and very cold winters
		w	Severe dry winters (subarctic)	a	Hot summer season
				b	Warm summer season
				c	Cool summer season
				d	Cool summers and very cold winters
E	**Polar** Average temperature of the warmest month <50°F/10°C	T	Warmest month's average temperature between 32°F/0°C and 50°F/10C		
		F	Warmest month's average temperature <32°F/0°C		

these groups, and in some cases further subdivisions describing seasonal variations within a climate set.

Climate may be thought of as the average weather in a given location. These general conditions vary over geographic space. Consider the differences between the hot wet tropics close to the Equator and the cold dry conditions typical of the Poles. While the Poles and the Equator illustrate extremes, for every place between them there exists an intermediate climate. Describing an area's average weather is difficult because atmospheric conditions, such as temperature and humidity, vary hourly, daily, seasonally, annually, and through longer periods. However, the climate zones may be identified by generalizations of several variables, based on long-range observations. *See* CLIMATOLOGY; DESERT; EQUATOR; NORTH POLE; SOUTH POLE.

Wladimir Köppen (1846–1940) was a German climatologist. Though weather data were incomplete at the time, Köppen, who was also a trained botanist, recognized that plant distributions were excellent proxies for physical conditions on land. With vegetation as a guide, he and his student/assistant, Rudolph Geiger, developed the first global climate classification map in 1928. The initial draft divided the climates into five major groups, based on inferred temperature and precipitation patterns. The zones were referred to as A (hot and humid), B (dry), C (midlatitude with mild winters), D (midlatitude with cold winters), and E (polar). *See* PLANT GEOGRAPHY; PLANTS, LIFE FORMS OF.

The first divisions in the hierarchy were quite broad and vague, encompassing many possible subclimates. Categories C, D, and E describe their geography more explicitly than they do their associated climate, since latitude and associated solar intensity are among the most significant controls of a location's climate. In addition, category B describes only the moisture domain without regard to temperature range, as it may refer to hot or cool deserts.

While the major headings are alphabetical, the second and third (if present) subdivisions are less intuitively labeled. Since the subsequent letters are abbreviations for German words, some of them are capitalized and some are not. This has proven confusing to some. Twenty-three originally specified climates are described in the **table**.

The second division denotes further detail, once again considering temperature and/or precipitation characteristics. For example, category B is subdivided into arid regions (deserts), W, and semiarid regions, S, known as steppes. Desert regions are therefore designated BW in the Köppen system. However, some second-order categories denote temperature ranges instead; for example, the polar climate group consists of ice cap (most frigid) and tundra (less cold) subclasses.

Some categories are further divided into subclasses, which are indicated by lowercase letters. For example, deserts (BW) may be further described as warm or cool; the Köppen codes for these climates are BWh and BWk, respectively.

Several changes were made by Köppen, and other geographers have modified the system since his death. A major category for highland, H, was added after the publication of the first matrix, bringing the total number of terrestrial climates to 24. The **illustration** shows a simplified map of several Köppen climate zones.

Classifying climate, or almost any other physical trend, is an attempt to organize natural phenomena. As a result, there are inherent inconsistencies and subjectivity in almost all classification systems, and Köppen's is no exception. Some argue that Köppen's system ignores several significant variables, such as global wind belts, air mass source regions, and frontal zones. Furthermore, the study of global climate change has supported a call for a quantitative approach to climate classification. Digital meteorological data collected globally via satellite can be analyzed by computers, yielding highly accurate, timely, and detailed maps. Even small alterations of a region's climate can be monitored in real time. However, the availability of superior meteorological monitoring equipment, including remote sensors, and nearly 80 additional years of archived weather since Köppen's death have resulted in the development of only a few alternative classification systems. Even so, the descriptive, yet empirical system devised by Köppen remains a model still in use. *See* GLOBAL CLIMATE CHANGE; METEOROLOGICAL SATELLITES; REMOTE SENSING.

Paula Messina

Bibliography. A. Getis, J. Getis, and J. D. Fellman, *Introduction to Geography*, 8th ed., 2001; A. M. Strahler, *Introduction to Physical Geography*, 1965; E. J. Tarbuck and F. K. Lutgens, *Earth Science*, 11th ed., 2005.

Kronig-Penney model

An idealized, one-dimensional model of a crystal which exhibits many of the basic features of the electronic structure of real crystals. Consider the potential energy $V(x)$ of an electron illustrated in **Fig. 1** with an infinite sequence of potential wells of depth $-V_o$ and width a, arranged with a spacing b. The Schrödinger wave equation can be readily solved for such an arrangement to give electron energy as a function of wave number.

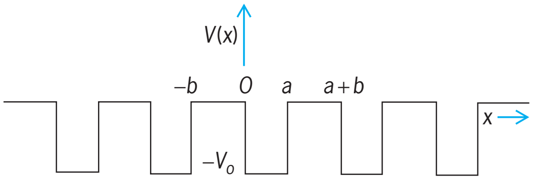


Fig. 1. Potential energy which is assumed for the one-dimensional Kronig-Penney model.

The energy bands that are thus obtained with the choice of constants given in Eq. (1) and (2) are

$$\left(\frac{2ma^2V_o}{\hbar^2}\right)^{1/2} = 12.0 \tag{1}$$

$$\frac{b}{a} = 0.1 \tag{2}$$

shown in **Fig. 2**. Notice that the width and the curvatures of the allowed bands increase with energy. The Kronig-Penney model has been extended to include the effects of impurity atoms. A solution can also be derived from the Dirac equation. *See* BAND

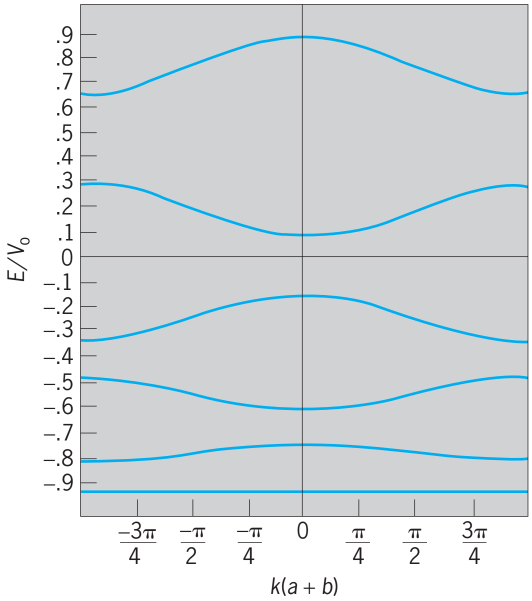


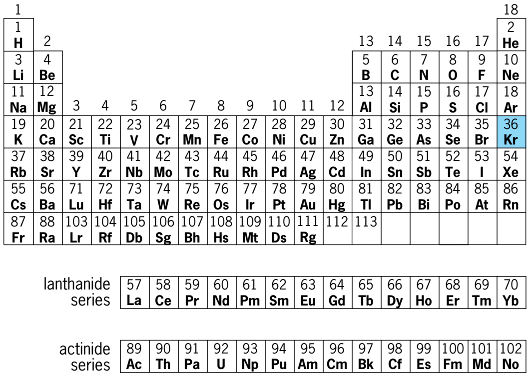
Fig. 2. The ratio of the electron energy *E* to the magnitude of the well depth *V*$_o$ shown as a function of electron wave number *k* times the "lattice constant" (*a* + *b*) for the Kronig-Penney model.

THEORY OF SOLIDS; NONRELATIVISTIC QUANTUM THEORY. Joseph Callaway

Bibliography. C. Kittel, *Introduction to Solid State Physics*, 7th ed., 1996; R. de L. Kronig and W. G. Penney, Quantum mechanics of electrons in crystal lattices, *Proc. Roy. Soc. (London), Ser. A*, 130:499–513, 1931; G. Wannier, *Elements of Solid State Theory*, 1959, reprint 1970.

Krypton

A gaseous chemical element, Kr, atomic number 36, and atomic weight 83.80. Krypton is one of the noble gases in group 18 of the periodic table. Krypton is a colorless, odorless, and tasteless gas. The **table** gives some physical properties of krypton. The principal use for krypton is in filling electric lamps and electronic devices of various types. Krypton-argon mixtures are widely used to fill fluorescent lamps. *See* INERT GASES; PERIODIC TABLE.



1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1 H																	2 He
3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
11 Na	12 Mg											13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
55 Cs	56 Ba	71 Lu	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
87 Fr	88 Ra	103 Lr	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112	113					

lanthanide series	57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb
actinide series	89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No

The only commercial source of stable krypton is the air, although traces of krypton are found in minerals and meteorites.

Physical properties of krypton

Property	Value
Atomic number	36
Atomic weight (atmospheric krypton only)	83.80
Melting point, triple point °C	−157.20
Boiling point at 1 atm pressure, °C	−153.35
Gas density at 0°C and 1 atm pressure, g/liter	3.749
Liquid density at its boiling point, g/ml	2.413
Solubility in water at 20°C, ml krypton (STP) per 1000 g water at 1 atm partial pressure krypton	59.4

A mixture of stable and radioactive isotopes of krypton is produced in nuclear reactors by the slow-neutron fission of uranium. It is estimated that about 2×10^{-8}% of the weight of the Earth is krypton. Krypton also occurs outside the Earth. Arthur W. Francis

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; D. R. Lide, *CRC Handbook Chemistry and Physics*, 85th ed., CRC Press, 2004; M. Ozima and F. A. Podosek, *Noble Gas Geochemistry*, 2001.

Kudzu

A perennial vine legume, capable of rapid growth in a warm temperate, humid subtropical climate. The name kudzu has a Japanese origin. Kudzu (*Pueraria thunbergiana*) was introduced into the United States in 1876 and used as a shade plant until 1906, when a few enthusiastic growers in the southeastern United States began to use it as a forage crop, a practice that continued for 30 years. It was then promoted as a soil-conserving plant. However, much prejudice developed against its use because of its spread into forest borders, drainage ditches, and other areas.

Kudzu is not adapted to tropical or arid climates or to alkaline soils, and probably requires winter cold for growth rejuvenation. The technique for growing kudzu successfully is to set out a few well-developed plants, use enough commercial fertilizer or equivalent for good growth, and protect the plants from weeds, insects, and grazing animals. The vines will spread over the area to be covered, producing roots at the nodes, which, unless killed by severe winter freezing, become independent plants. Thus, a few plants produce many in 2 or 3 years if adequately protected. *See* FERTILIZING.

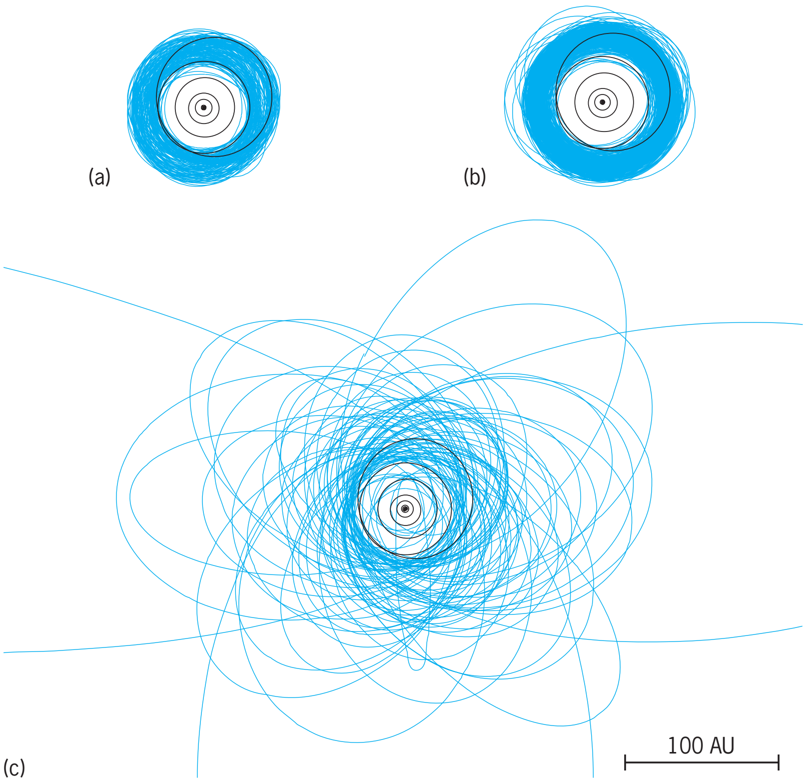
Kudzu produces moderate yields of forage, but it must be grazed with care to prevent loss of stand. Since the viney stems are difficult to harvest for hay, adapted grasses and clovers are preferred as forage crops.

A simple, economical method is needed for keeping kudzu from spreading into areas where it is not wanted. It may be stopped by a permanent pasture at the border of a kudzu field, or by repeated cultivation or harrowing the borders of cropland. No acceptable method of spread prevention has been developed for uncultivated or nonpastured areas.

Tropical kudzu (*P. phaseolides*) is one of the most important and widely planted cover and green manure crops of the tropics. It makes rapid vigorous growth, providing quick ground cover and suppressing most other vegetative growth. It is used as a forage crop although careful management is required to prevent complete domination of mixtures with grasses and other species. Its habit of growth is similar to kudzu of subtropical areas. *See* LEGUME FORAGES; ROSALES. Paul Tabor

Kuiper Belt

A vast reservoir of icy bodies in the region of the solar system beyond the orbit of the planet Neptune. The trans-Neptunian region of the solar system was thought to be home only to the small planet Pluto and its satellite Charon until the discovery of the Kuiper



Plane views of the outer solar system showing (*a*) resonant Kuiper Belt objects, (*b*) classical KBOs, and (*c*) scattered KBOs. The types of KBOs are represented by a band of colored orbits in each diagram, while the black concentric circles in the center of each diagram are the orbits of the giant planets, and the black ellipse is the orbit of Pluto. (*Ronnie Hoogerwerf, Smithsonian Astrophysical Observatory*)

Belt object 1992 QB_1 in 1992 by David Jewitt and Jane Luu. Members of the belt (Kuiper Belt objects, or KBOs) are part of the original building blocks (planetesimals) of the outer planets, left behind simply because they were not incorporated into planets or ejected from the solar system. They are relics from the solar nebula, the original disk of gas and dust that gave rise to everything in the solar system. As such, Kuiper Belt objects should preserve records of the early stages of the solar system's evolution. *See* NEPTUNE; PLANET; PLUTO.

Prediction and discovery. The idea that distant planetesimals that did not agglomerate into planets might still exist beyond the planets was proposed in 1951 by Gerard Kuiper. However, Kuiper had a little-known predecessor: the concept of a trans-Neptunian belt of icy bodies was first mentioned by K. E. Edgeworth in 1949. In the 1980s the concept of the Kuiper Belt was revived as a possible source for the short-period comets (those which take less than 200 years to complete one orbit around the Sun). Long-period comets, whose orbital periods are longer than 200 years, have long been recognized to come from the Oort Cloud, a large spherical cloud of comets at roughly 10,000 astronomical units (AU) from the Sun. (An astronomical unit is equal to the distance between the Earth and the Sun, about 1.5×10^8 km or 9.3×10^7 mi.) However, the Oort Cloud could not be the source for the short-period comets, and theoretical arguments suggested that the short-period comets likely originated in a comet belt beyond Neptune. *See* COMET.

Edgeworth and Kuiper had no theoretical basis for their bold idea. They simply thought that there was no a priori reason why the solar system should end abruptly at Neptune or Pluto. Perhaps large planets could not form beyond Neptune, but smaller bodies might be able to. Edgeworth and Kuiper envisioned a population consisting of residual planetesimals left over after planets had finished their formation. Because of the low temperatures in the trans-Neptunian area, these distant objects were likely to be ice-rich, identical to comets. At the time of Edgeworth's and Kuiper's hypothesis, Pluto was the only permanent solar system object known beyond Neptune, and the apparent emptiness of the outer solar system was rather startling compared to the rich abundance of bodies in the inner solar system. In 1987 Jewitt and Luu started a telescopic survey to test this apparent emptiness of the outer solar system. The search finally succeeded in August 1992, when 1992 QB_1 was found.

Population. The Kuiper Belt begins just beyond Neptune and extends to some as yet unknown outer limit. KBOs are not distributed uniformly throughout the belt, but lie in dense bands separated by empty space (see **illustration**). The bands represent stable regions where bodies can survive for the age of the solar system, while empty regions represent unstable areas where previous KBOs have been cleared out by the gravitational influence of Neptune. Based on how their orbits are distributed throughout the region, KBOs are classified as resonant, classical, or scattered.

Resonant KBOs. The Kuiper Belt contains several special areas called mean motion resonances, which are best explained by an example: Located at 39.5 AU, Pluto lies in the 3:2 resonance with Neptune. This means that every time Neptune completes three orbits around the Sun, Pluto completes two such orbits. In the same vein, objects in the 7:4 resonance complete 4 orbits for every 7 orbits completed by Neptune, objects in the 8:5 complete 5 orbits for Neptune's 8 orbits, and so forth. These special relationships with Neptune protect the resonance objects from close encounters with the planet, and the resonance orbits are therefore long-lived. Resonances thus act as stable oases in space where members can survive for the age of the solar system. Just as the 3:2 resonance protects Pluto, it also protects other objects in this resonance (called plutinos to highlight their dynamical similarity with Pluto). The current estimate is that the resonant KBOs (with the plutinos dominating in number) make up roughly 20% of the Kuiper Belt population (illus. *a*), with Pluto being the largest known Kuiper Belt object.

Classical KBOs. The classical KBOs are characterized by near-circular orbits that lie almost entirely in the ecliptic plane between 42 and 48 AU (outside the resonances) and make up about 45% of the total population (illus. *b*). These orbits are what would be expected from the first-generation planetesimals in the solar nebula, suggesting that the classical KBOs are indeed primordial planetesimals that have managed to preserve their original orbits. Unexpectedly, the classical Kuiper Belt possesses a well-defined edge at approximately 50 AU: as of late 2006, all known KBOs beyond this edge belong to the scattered Kuiper Belt population (see below). The origin for this edge is still not clear: the proposed hypotheses generally postulate that the classical Kuiper Belt beyond 50 AU had been stripped by scattering processes, such as the passage of Earth-sized bodies through the belt or a close encounter with a star. The competing hypotheses make testable predictions, and astronomers hope that as the structure of the Kuiper Belt is better defined, it will help narrow down the competing hypotheses.

Scattered KBOs. The scattered KBOs make up the remaining 35% of the Kuiper Belt, and they easily stand out from the rest of the belt with their very large, very elliptical orbits (illus. *c*). Whereas the known classical and resonant KBOs are clustered inside 50 AU, the object 2000 OO_{67}, the most distant scattered Kuiper Belt object yet discovered, reaches at least to 1053 AU at its farthest point from the Sun. Several theories have been proposed to explain the origin of the scattered KBOs. One theory postulates that the belt is a by-product of the scattering process that produced the Oort Cloud. As Uranus and Neptune reached their full sizes, they scattered nearby comets to very large distances, forming the Oort Cloud; those comets scattered to less extreme distances would have formed the scattered Kuiper Belt population.

Total population. The current total mass of the Kuiper Belt population between 30 and 50 AU is estimated to be roughly 0.1 earth mass, 100 times more massive than the asteroid belt. About 1100 KBOs were known in August 2006 with diameters estimated in the 100–1000-km (60–600-mi) range. (In comparison, Uranus and Neptune are about 50,000 km or 30,000 mi in diameter, while Pluto has a diameter of 2320 km or 1440 mi.) There are currently about 30,000 KBOs with diameters larger than 100 km (60 mi) in the entire Belt, making them roughly 100 times as numerous as similarly sized bodies in the asteroid belt.

Large KBOs. Objects smaller than 100 km are numerous but too small to be detected with conventional telescopes; on the other hand, bodies with diameters of 1000 km (600 mi) or more are easily detectable from the ground but are relatively rare. Surveys have been undertaken that are aimed at both large and small KBOs. By August 2006, roughly 10 KBOs in the 1000-km-diameter range had been discovered by several surveys specifically aimed at the largest bodies. An example of such a survey is the Palomar survey carried out by the California Institute of Technology. Like all other surveys aimed at KBOs, its strategy is to search as much of the sky as possible, at the expense of sensitivity; this strategy works because large KBOs are bright but very rare, so the key to finding them is large sky coverage. These 1000-km objects bridge the gap between the relatively abundant smaller (100 km diameter) bodies and Pluto (2320 km diameter), hitherto the largest Kuiper Belt object known. However, other bodies may have been found that would strip Pluto of such distinction. A case in point is the KBO 2003 UB_{313}, named v36 199 Eris in September 2006, which soon after its discovery was postulated to be even larger than Pluto based on its absolute magnitude. The absolute magnitude is the brightness of an object corrected to a standard distance from the Sun and the Earth, and is an indication of an object's size. However, since the absolute magnitude is equal to the albedo (reflectivity) times the squared diameter, an absolute magnitude brighter than Pluto's is not sufficient to indicate unambiguously a larger size. If the albedo were known, the diameter could be disentangled from the absolute magnitude. The albedo is usually determined by observing an object in the thermal infrared, measuring the object's heat radiation. Such measurements are difficult for KBOs because, due to their great distance from the Sun and cold temperatures, their thermal radiation peaks near 60–70 micrometers, where the Earth's atmosphere is opaque. Until space-based telescopes (such as, NASA's *Spitzer Space Telescope*) can deliver thermal measurements of a large sample of KBOs, the albedos of the largest KBOs can be determined via observations in the submillimeter. As of August 2006, the true size of Eris was still uncertain. Observations by the *Hubble Space Telescope* suggested that Eris was roughly 5% larger than Pluto, but measurement uncertainties prevent a definitive assessment. Whether Eris is larger than Pluto remains to be confirmed by future measurements. *See* ALBEDO; INFRARED ASTRONOMY; SPITZER SPACE TELESCOPE; SUBMILLIMETER ASTRONOMY.

The discovery of each large KBO is often greeted with great excitement as it raises the hope of adding yet another planet to the solar system's inventory, but in truth such large objects are expected if, as is widely believed, KBOs formed basically in the same manner as other bodies in the solar system, that is, by accretion from smaller bodies (planetesimals). According to current accretion models, about 100 such bodies are expected. Besides their appeal to the general public, large KBOs are important for several scientific reasons:

(1) They are bright enough for detailed investigations of their physical properties. The more common 100-km bodies are too faint for anything other than the most rudimentary measurements (such as optical brightness through various filters), and offer little information on their physical makeup. Barring close-up visits with spacecraft in the distant future, astronomers could only hope to extract meaningful physical information such as surface composition, shape, and spin rate from the large bright KBOs.

(2) The number of these large bodies, compared to the number of smaller bodies, is a test of the existing accretion models. These models predict the relative numbers of large and small KBOs, and their confirmation (or the lack thereof) has significant implications for the conditions in the solar nebula when the KBOs were forming.

(3) The largest KBOs seem to be more reflective

than the smaller ones. This suggests that large KBOs may be better at retaining icy material than their smaller counterparts, much like Pluto (albedo = 0.6).

90377 Sedna. In 2004 the Kuiper Belt object 90377 Sedna was discovered in the Palomar survey. Sedna is of great interest for two reasons:

(1) Sedna's perihelion distance (closest approach to the Sun) is 76 AU, well beyond the scattering influence of Neptune. In this sense it resembles the Kuiper Belt object 2000 CT_{105} (perihelion distance 44 AU). It is not clear how these two KBOs attained their orbits beyond Neptune. Their existence seems to suggest an early epoch of strong scattering events during which smaller (less than planet size) solar system bodies were scattered by a configuration of planet-size bodies different from what is known today.

(2) Sedna's large size (1000–1500 km or 600–900 mi) means that it could not have formed in its current orbit far beyond the planets; there is little doubt that Sedna must have formed closer in and was subsequently scattered outward. Sedna's classification is still uncertain. Since it lies between the known Kuiper Belt and the classical Oort Cloud, it has been suggested to be part of the Inner Oort Cloud. This inner part of the Oort Cloud may connect the classical Oort Cloud and the Kuiper Belt.

Small KBOs. Although small KBOs do not tend to generate as much excitement as large ones, there is also much interest in how to detect the small end of the size distribution (KBOs with kilometer-size diameter or smaller). These very small bodies are too faint to be detected by conventional imaging methods and are generally expected to be "seen" only by the occultation method. With this method, the presence of these bodies can be inferred when the bodies pass in front of stars and momentarily block out the stars' light (they occult the stars). The most concerted effort to detect KBOs via occultation is the Taiwan American Occultation Survey (TAOS), which is aimed at detecting KBOs a few kilometers in diameter. The survey's four 0.5-m (20-in.) robotic telescopes will monitor 3000 stars every clear night for several years to search for these occultation events and will send out alerts when events occur. Follow-up observations can then be carried out with larger telescopes. However, there is already a tantalizing result on the smallest end of the Kuiper Belt population: in 2006 a group of Taiwanese astronomers reported occultations in x-rays from Scorpius X-1, the brightest source of x-ray light in the sky, and attributed these occultations to 10–100-m bodies in the Kuiper Belt. However, the occultation rate implies 10^{15} bodies in the 10–100-m range, a population much larger than that predicted by models (10^9–10^{12}). This result remains to be confirmed either by another occultation survey or by space-based missions that can actually monitor the diffraction patterns of the occultation events and establish the distance of the occulting bodies. *See* OCCULTATION.

Population loss. The Kuiper Belt has been much eroded since its formation. The original Kuiper Belt is estimated to have contained roughly 10 earth masses, or 100 times as much mass as the present Kuiper Belt.

Physical properties. The properties of KBOs are of intense interest since these bodies are primordial survivors from the early solar nebula. Unfortunately, most KBOs are so faint that detailed investigations of their properties are extremely difficult. The best determined property is probably their surface colors, which turn out to cover a broad range from neutral to very red. This may be explained in the context of their composition, which is believed to be rich in molecular ices that are abundant in the solar system (water, ammonia, and methane). Laboratory experiments lead to the prediction that after prolonged irradiation by high-energy particles such as cosmic rays and solar-wind particles, KBOs should be covered by an irradiation mantle, a carbon-rich crust that covers the ices buried underneath. Occasional collisions among KBOs may puncture the irradiation crust, revealing pristine unirradiated material which would be of a different color from the crust. The degree to which a Kuiper Belt object is covered by craters and impact debris may explain the wide range of colors.

Given the difficulties in determining the surface properties of KBOs, astronomers look to brighter (larger) analogs of KBOs for clues. A popular counterpart of the KBOs is Triton, a 2700-km-diameter (1680-mi) satellite of Neptune; it orbits Neptune "backwards" (in retrograde orbit) and therefore is suspected to be a captured Kuiper Belt object. Pluto, the largest Kuiper Belt object known, is also often used as a template for KBOs. However, the atmospheres of both Triton and Pluto complicate comparisons with the average Kuiper Belt object with regard to surface properties such as spectra and albedos.

The Kuiper Belt also contains other binary systems besides Pluto and its satellite Charon: several KBOs are now known to have their own moons. Like Charon, these satellites of KBOs are large compared to the primary (as much as one-third the size), but unlike Charon they are also quite far (10,000–100,000 km or 6000–60,000 mi) from the primary. Their formation mechanism, like so many other aspects of the Kuiper Belt, remains a mystery. Binary formation is strongly affected by the dynamical environment in which it occurs, so once understood, binary formation promises to yield important clues to the early Kuiper Belt environment. *See* BINARY ASTEROID; SOLAR SYSTEM.

Jane Luu

Bibliography. D. Jewitt and J. Luu, Discovery of the candidate Kuiper Belt object 1992 QB_1, *Nature*, 362:730–732, 1993; G. P. Kuiper, Planets and satellites, in G. P. Kuiper and B. M. Middlehurst (eds.), *The Solar System*, University of Chicago Press, 1961; J. Luu and D. Jewitt, Kuiper Belt objects: Relics from the accretion disk of the Sun, *Annu. Rev. Astron. Astrophys.*, 40:63–101, 2002.

Kumquat

Shrubs or small trees that are members of the genus *Fortunella*, which is one of the six genera in the group of true citrus fruits. Kumquats are believed to

Fruit cluster and foliage of Nagami kumquat (*Fortunella margarita*). (*McFarland Co.*)

have originated in China and the Malay Peninsula, but are now widely grown in all citrus areas of the world. Of the several species the most common are *F. margarita*, which has oval-shaped fruit (see **illus.**), and *F. japonica*, which has round fruit. *See* SAPINDALES.

The kumquat's stems, leaves, flowers, and fruits resemble those of *Citrus*. Kumquats bear numerous flame- to orange-colored fruits of small size, often less than 1 in. (2.5 cm) in diameter, having three to five locules filled with an acid pulp and a sweet, edible pulpy rind. The trees are among the most cold-hardy of the citrus fruits; they stay dormant even during protracted warm spells in the winter months, which enables them to withstand low temperatures.

Kumquats, with their brilliant orange-colored fruits and dense green foliage, are highly ornamental and are most frequently grown for this reason. Sprays of foliage and fruit are commonly used for decoration, particularly in gift packages of ordinary citrus fruits. Kumquat fruits can be eaten whole without peeling; they are also used in marmalades and preserves and as candied fruits. *See* FRUIT, TREE.

Frank E. Gardner

Kuroshio

A swift, intense current flowing northeastward off the coasts of China and Japan in the upper waters of the North Pacific Ocean. The Kuroshio is the western portion of a giant clockwise, horizontal circulation known as the North Pacific subtropical gyre (**Fig. 1**). This circulation extends from 15° to 45°N across the entire width of the Pacific Ocean. It is driven by the large-scale winds—the trades in the south and the westerlies in the north. As with all other western boundary currents, such as the Gulf Stream, the effect of the Earth's rotation and its spherical shape is to concentrate the Kuroshio flow into a current that is only about 100 km (50 nmi) wide with speeds up to 2 m/s (4 mi/h). *See* CORIOLIS ACCELERATION; GULF STREAM; PACIFIC OCEAN.

The Kuroshio (Japanese, meaning "Black Current") has an apparent blackness resulting from the water clarity, which is a consequence of the low biological productivity of seawater in the area. It originates off the southeast coast of Luzon, the main island of the Philippines. For the first 1000 km (500 nmi), the Kuroshio flows northward along the east coasts of Luzon and Taiwan, until it enters the East China Sea. For the next 1000 km, it flows northeastward near the edge of the continental shelf off eastern China, until it exits the East China Sea through the Tokara Strait. During its final 1000 km, it flows east-northeastward off the southern coast of Japan (where it is sometimes called the Japan Current). Finally, it leaves the Asian coast near Tokyo and travels into the interior of the North Pacific Ocean as a slowly expanding jetlike current known as the Kuroshio Extension (**Fig. 2**). Here it merges with the Oyashio, a cold current with high biological productivity, and becomes the North Pacific Current (Fig. 1).

Like the Gulf Stream in the North Atlantic, the Kuroshio rapidly carries large quantities of warm water from the tropics into midlatitude regions. It is consequently an important agent in redistributing global heat. North of 30°N, where prevailing winds are westerlies, the North American climate is strongly affected by the warmth of these waters. Kuroshio volume transport is huge (more than 100 times that of the Amazon River): typically $2\text{–}3 \times 10^7\ \text{m}^3/\text{s}$ ($7\text{–}10 \times 10^8\ \text{ft}^3/\text{s}$) until leaving the East China Sea, and then growing to $6 \times 10^7\ \text{m}^3/\text{s}$ ($20 \times 10^8\ \text{ft}^3/\text{s}$) south of Japan. One reason for the increase is the merging of the Kuroshio flow from the East China Sea with northward subtropical gyre flow traveling outside this sea, along the eastern edge of the Ryukyu island chain. Another reason is the contribution of the flow occurring in a local recirculating gyre (Fig. 2).

The Kuroshio and Kuroshio Extension (like the Gulf Stream) are thought to be flanked by a number of such recirculating gyres, which are typically several hundred kilometers across. According to theory, these small gyres may be caused by eddying motions, which are a result of the Kuroshio flow becoming unstable. The instability first appears as side-to-side motions or "meanders" of the Kuroshio current core. The meanders are wavelike in character and usually

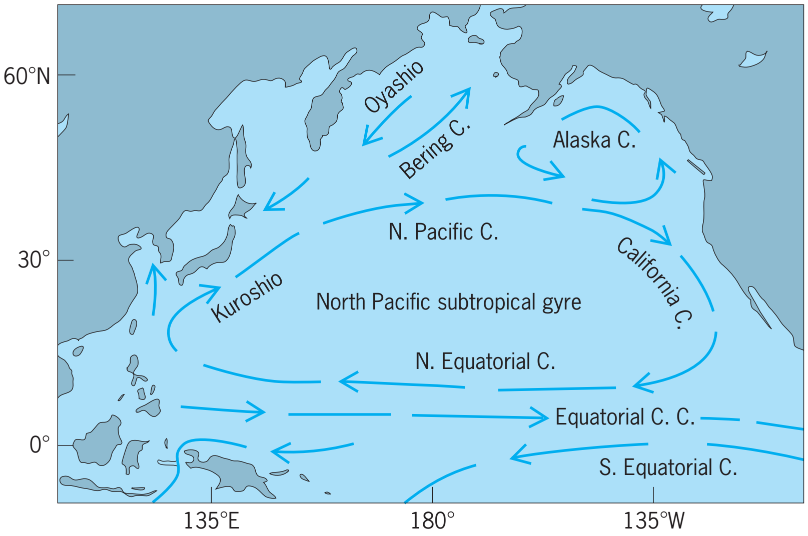


Fig. 1. General surface circulation of the North Pacific Ocean. (*Adapted from Fig. 2.17 in J. R. Apfel, Principles of Ocean Physics, 1987, Academic Press*)

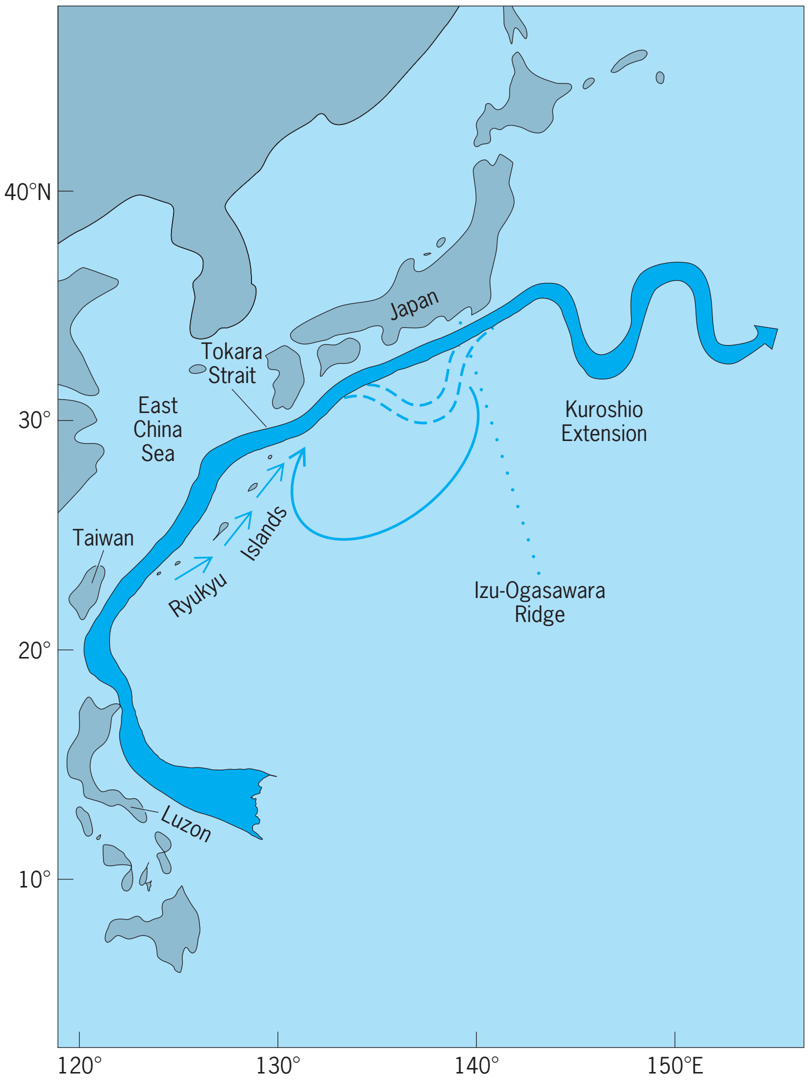


Fig. 2. Flow path of the Kurishio and Kurishio Extension. Broken lines show the alternative "large-meander" path south of Japan. Smaller arrows show other flows described in the text. (*After Y. G. Guo, The Kurishio, Part II, Oceanog. Mar. Bio. Annu. Rev., 29:155–189, 1991*)

propagate slowly (<25 km/day or 0.5 nmi/h) downstream. Occasionally, especially in the Extension region, such a meander can grow until it pinches off into a separate, spinning ring of water about 200 km (100 nmi) in diameter, which can survive for many seasons.

One unique form is the so-called Kuroshio large meander. This stationary meander is typically a 300-km (180-mi) southward loop which forms intermittently off southeast Japan. At present, it occurs about one-third of the time, and once formed lasts 2 to 10 years. It takes place just upstream of where the Kuroshio crosses the Izu-Ogasawara Ridge, and the ridge appears to play a role in causing the meander (Fig. 2). There is evidence that the large meander is associated with weak subtropical gyre flow resulting from weakening of atmospheric high pressure over the subtropical North Pacific. Proxy indicators, in dated sediments from the northern Izu-Ogasawara Ridge, have shown how the tendency to form a large meander there has changed over time, with a pronounced minimum 7000 years ago. These data suggest a strong atmospheric subtropical high at that time, with resulting swift winds and intense gyre circulation. *See* PALEOCLIMATOLOGY.

The Kuroshio is readily apparent in data obtained from satellites. In particular, warm surface water carried by the Kuroshio shows up clearly in sea-surface temperature images. Also, satellite altimeters, which are today able to detect sea-surface height changes of a few centimeters, can readily measure the change associated with the Kuroshio: a phenomenon known as geostrophy, resulting from the rotation of the Earth, makes the offshore side of the Kuroshio about 1 m (3 ft) higher than the inshore side. *See* GEOSTROPHIC WIND; REMOTE SENSING.

Mark Wimbush

Bibliography. Y. J. Guo, The Kuroshio, Part II: Primary productivity and phytoplankton, *Oceanog. Mar. Biol. Annu. Rev.*, 29:155–189, 1991; Y. Hirota, The Kuroshio, Part III: Zooplankton, *Oceanog. Mar. Biol. Annu. Rev.*, 33:151–220, 1995; N. G. Hogg and W. E. Johns, Western boundary currents, *Rev. Geophys.*, 33(suppl.):1311–1334, 1995; H. Kawai, A brief history of recognition of the Kuroshio, *Prog. Oceanog.*, 41:505–578, 1998; K. Sawada and N. Handa, Variability of the path of the Kuroshio ocean current over the past 25,000 years, *Nature*, 392:592–595, 1998; H. Stommel and K. Yoshida (eds.) *Kuroshio: Physical Aspects of the Japan Current*, University of Washington Press, 1972; J. L. Su, B. X. Guan, and J. Z. Hang, The Kuroshio, Part I: Physical features, *Oceanog. Mar. Biol. Annu. Rev.*, 28:11–71, 1990.

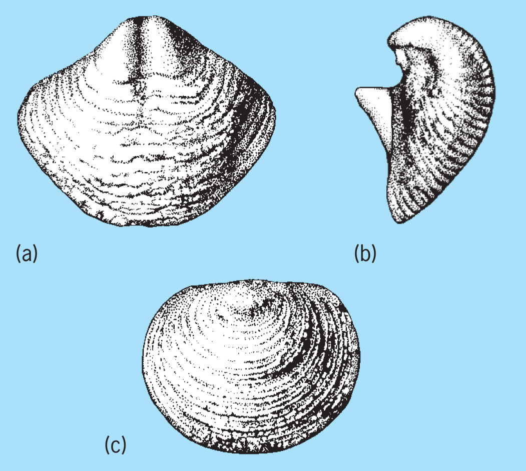
Kutorginata

A small extinct class and one of the most primitive types of rhynchonelliformean brachiopods. The kutorginates were widely distributed from the Early Cambrian and became extinct at the end of the Middle Cambrian. *See* BRACHIOPODA; RHYNCHONELLIFORMEA.

Classification. The Kutorginata appear in the classification scheme as follows:

- Phylum Brachiopoda
  - Subphylum Rhynchonelliformea [Articulata]
    - Class Kutorginata
      - Order Kutorginida
        - Superfamily Kutorginoidea
          - Family Kutorginidae
            - 5 genera (Early to Middle Cambrian)
  - Superfamily Nisusioidea
    - Family Nisusiidae
      - 5 genera (Early to Middle Cambrian)

Morphology. Kutorginates have a biconvex, fibrous, impunctuate (lacking holes) shell (see **illustration**). On the posterior margin of the shell, the pedicle (a tough but flexible stalklike appendage used for attachment) protrudes through a large median opening (delthyrium and notothyrium) located between the valves. In the ventral valve



***Kutorgina*, showing (*a*) exterior of pedicle (ventral) valve; (*b*) lateral view of shell; (*c*) exterior of brachial (dorsal) valve. (*After C. D. Walcott, Cambrian Brachiopoda, USGS Monogr., vol. 51, 1912*)**

(formerly named pedicle valve), this opening is partly covered by a triangular convex plate (pseudodeltidium) bounded laterally by furrows, acting in the shell articulation. The pedicle appears at least as strong as that of the coeval (contemporary) lingulids. The putative coelomic cavity is similar anatomically to those of the Cambrian linguliformeans. *See* LINGULIDA.

The articulation of the shell consists of edges of the interareas without teeth and dental sockets: the valves rotate on simple hinge mechanisms that are different from those of other rhynchonelliformean brachiopods. A small apical foramen (opening) may occur on the ventral valve.

The visceral cavity is in a posteromedian position and extends anteriorly over one-third of the shell length. The muscle arrangement is very similar to that of other rhynchonelliformean brachiopods with adductor and diductor muscle systems (that is, muscles used to close and open the shell, respectively). The complete alimentary canal consists of the mouth located at the base of the lophophore (a food-gathering and respiratory organ), a posterodorsal esophagus, an enlarged pouch representing the stomach, a narrow intestine, and a dorsoposterior anus located near the proximal end of the pedicle. The adult lophophore development attains an early spirolophe (a lophophore with spiraled appendages) stage.

The main mantle canals known as vascula lateralia are exclusively disposed radially (pinnate type) in both valves. On the dorsal side, the vascula media are bifurcate.

Members of this group were sessile, but the epifaunal (surface-attaching) position remains questionable. Christian C. Emig

Bibliography. R. L. Kaesler (ed.), *Treatise on Invertebrate Paleontology*, Part H, revised, *Brachiopoda* Geological Society of America and University of Kansas, 2000; L. E. Popov, A. J. Rowell, and J. S. Peel, Early Cambrian brachiopods from North Greenland, *Palaeontology*, 40:337–354, 1997; A. J. Rowell and N. E. Caruso, The evolutionary significance of *Nisusia sulcata*, an early articulate brachiopod, *J. Paleontol.* 59:1227–1242, 1985; C. B. Skovsted and L. E. Holmer, Early Cambrian brachiopods from Northeast Greenland, *Paleontology* 48(2):325–345, 2005; Z. Zhang et al., Rhynchonelliformean brachiopods with soft-tissue preservation from the Early Cambrian Chengjiang Lagerstätte of South China, *Plaeontology*, in press, 2006.

L

Labradorite — Lever

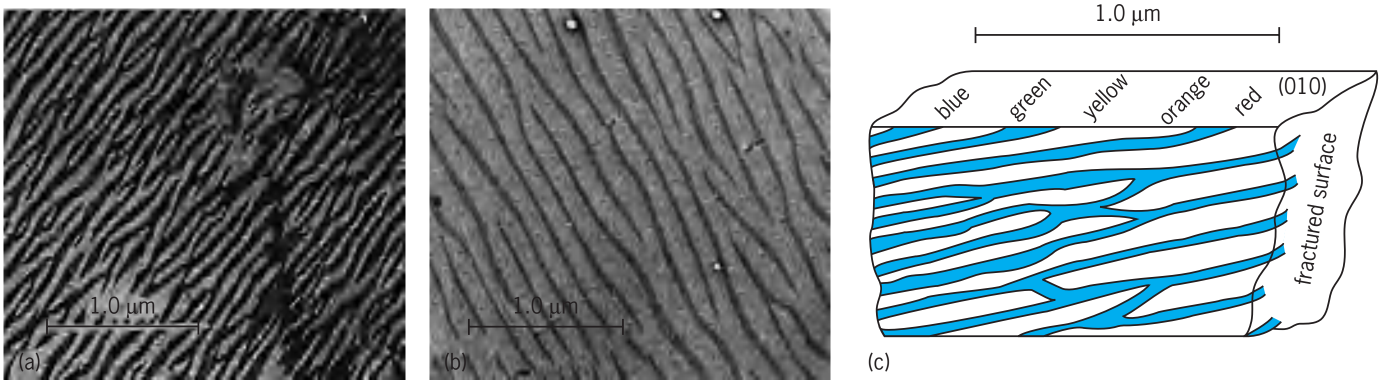
Labradorite

A plagioclase feldspar with composition $Ab_{50}An_{50}$ to $Ab_{30}An_{70}$ ($Ab = NaAlSi_3O_8$; $An = CaAl_2Si_2O_8$), occurring in basic igneous rocks, that is, basalts, gabbros, and anorthosites. Its hardness is 6–6.5; mean refractive index, 1.56; and specific gravity, 2.70. Labradorite crystals are translucent to dark or cloudy (from inclusions) but often display brilliant colors like those in peristerites and Huttenlocher intergrowths. In nature, most plagioclases throughout the composition range $Ab_{54}An_{46}$ to $Ab_{40}An_{60}$ are exsolved into two e-plagioclases that have compositions $Ab_{56\pm4}An_{44\mp4}$ and $Ab_{42\pm6}An_{58\mp6}$ and are coherently intergrown as somewhat irregular lamellar slabs 50 to 20 nanometers thick and known as Bøggild intergrowths (see **illus.**). Intense colors may result from constructive interference of selected wavelengths of light reflected at successive interfaces between the alternating Ab- and An-rich lamellae. In fact, interference colors are so commonly observed in this composition range that the phenomenon has been called labradorescence. Gemologists often use the nonmineralogical term spectrolite in describing them. Iridescent colors in hummingbird feathers and the wings of *Morpho* butterflies are physically similar phenomena. *See* ALBITE; ANDESINE; ANORTHOSITE; BASALT; FELDSPAR; GABBRO; IGNEOUS ROCKS.

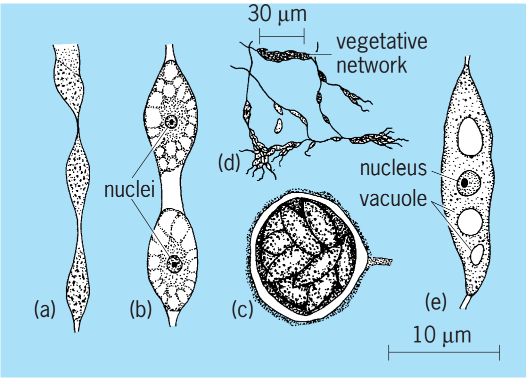
Paul H. Ribbe

Labyrinthulia

Protozoa forming a subclass of Rhizopodea with obscure relationships to rest of the class. There is one order, Labyrinthulida (see **illus.**). The mostly marine, ovoid to spindle-shaped, uninucleate organisms secrete a network of filaments (slime tubes) along which they glide, usually singly at rates of 4–150 micrometers/min. This network inspired the name "net slime molds" sometimes applied to them. The mechanism of locomotion is unknown. The unique *Labyrinthula minuta* may move for about 20 μm, then stop, and later move again, either



Photoemission electron microscope images of labradorite from an anorthosite in Finland. The more calcic phase is the lighter of the two. The apparent lamellar thickness are greatly exaggerated because the specimens are cut at ∼ 10° to the plane of the lamellae: (*a*) "blue" area and (*b*) "red" area (*courtesy of F. Laves*). (*c*) Schematic showing the approximate relations of lamellar thickness and interference color, as observed on the (010) cleavage surface. Bulk composition of the labradorite changes from about $Ab_{50}An_{50}$ (blue) to $Ab_{44}An_{56}$ (red).



***Labyrinthula zopfi*: (*a*) portion of living network, (*b*) two organisms stained, (*c*) encysted stage. *Labyrinthula macrocystis*: (*d*) vegetative network, (*e*) single organism stained. (*After R. P. Hall*, *Protozoology*, *Prentice-Hall*, *1953*)**

forward or backward. The slime filaments, which show an amorphous matrix, may or may not be tubular within a single culture. Individual organisms may move within or along the outside of a slime tube and occasionally may leave the track. A thin pellicle permits changes in diameter, but pseudopodia and phagotrophy have not been seen.

Fission has been described and, in certain species, formation of a sorus by aggregation of individuals. Some produce a pseudosorus by enlargement of one organism; inside, uninucleate pseudospores are produced. The organisms occur on algae and eelgrass, and may cause a disease of the latter. *See* PROTOZOA; RHIZOPODEA; SARCODINA; SARCOMASTIGOPHORA. Richard P. Hall

Bibliography. S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982; K. L. Pokorny, *J. Protozool.*, 14:697–708, 1967.

Lacquer

A fast-drying, hard, high-gloss surface coating. Lacquers are made by dissolving a cellulose derivative and other modifying materials in a solvent and adding pigment if desired.

The cellulose derivative most commonly used is nitrocellulose, but a cellulose ester such as cellulose acetate or cellulose butyrate, or a cellulose ether such as ethyl cellulose is often used in formulation. Nitrocellulose is not soluble in conventional paint thinners, and so a mixture of solvents with high volatility and extremely fast drying times is used. The mixture usually contains esters (ethylacetate), aromatic hydrocarbons (toluene), and petroleum thinners. Solvent mixtures are formulated carefully in order to produce the combination of solvency and evaporation rate which is required.

Nitrocellulose is extremely hard. Its flexibility is enhanced by addition in the formulation of plasticizers such as vegetable oils (castor or linseed) modified for compatibility with nitrocellulose, or chemical compounds such as dibutyl phthalate or tricresyl phosphate.

Adhesion of lacquers can be improved by the addition of other resins, and these also may reduce the cost by permitting the use of less expensive solvents or, alternatively, by allowing the application of heavier films.

Lacquers dry by evaporation of the solvent. They usually are applied by spray because of their rapid drying properties. The drying time may be extended by altering the formulation to allow application by brushing, but brushing lacquers seldom are used.

Lacquers have been used extensively as fast-drying, weather-resistant finishes for automobiles and as coatings for furniture and other factory-finished items. Their use has diminished, however, as less expensive coating materials with improved properties have been developed. *See* CELLULOSE; SURFACE COATING. C. R. Martinson; C. W. Sisler

Bibliography. *Kirk-Othmer Encyclopedia of Chemical Technology*, 4th ed., 1994.

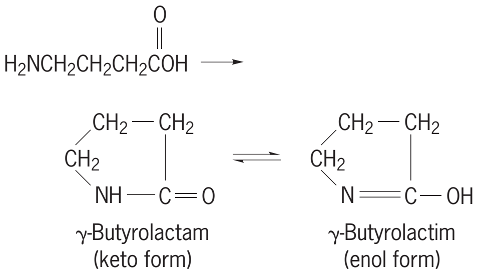
Lacrimal gland

A tubuloalveolar or acinous skin gland, also known as the tear gland. The lacrimal glands develop from the skin epithelium which folds inward over the developing eye. Two types occur among the vertebrates, the lacrimal proper and the Harderian. These eye glands are first found in the amphibians, associated with the inside of the lower eyelid. In urodele amphibians the eye gland extends along the inner aspect of the lower eyelid. In *Salamandra* it becomes divisible into an anterior Harderian gland associated with lower eyelid structures and a posterior lacrimal gland below the upper eyelid. In frogs and toads only the Harderian gland is present and is associated with the nictitating membrane or third eyelid which develops in relation to the lower lid.

The eye glands drain into the nasal cavity by means of the lacrimal duct. In higher vertebrates, with the exception of snakes, certain lizards, and geckos, Harderian and lacrimal glands are present. In the higher vertebrates the Harderian gland functions in relation to the third eyelid or nictitating membrane, and the lacrimal gland becomes located dorsally above the eyeball near the outer angle of the palpebral fissure. In aquatic mammals, such as seals, whales, and sea cows, the Harderian gland is the more prominent of the two sets of eye glands and it secretes a sebaceous or oily substance. In land mammals the lacrimal gland proper is highly developed as a complex tubuloalveolar structure with several ducts which pour their copious fluid into the outer, upper part of the conjunctival sac or cavity. The tear substance washes across the eyeball to the inner palpebral fissure or commissure. Eventually it passes through two small openings, one on the margin of each lid, into the lacrimal ducts. The latter converge to form the lacrimal sac, from which the nasolacrimal duct leads into the nasal passageway. Tears contain a considerable quantity of the common salt, sodium chloride. *See* EPITHELIUM; EYE (VERTEBRATE); EYE DISORDERS; GLAND. Olin E. Nelsen

Lactam

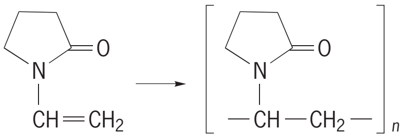
A cyclic amide that is the nitrogen analog of a lactone. For example, γ-aminobutyric acid readily forms γ-butyrolactam (also known as 2-pyrrolidinone) upon heating, as in reaction (1). The tautomeric enol form of a lactam is known as a lactim.



(1)

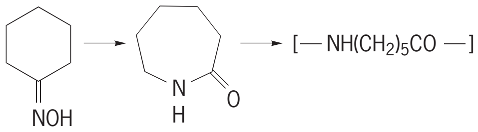
The δ-amino acids similary form δ (six-membered-ring) lactams upon heating, but larger- and smaller-ring lactams must be made by indirect methods, such as addition of ketenes to imines for β-lactams and ring expansion via Beckmann rearrangement of cyclic ketoximes for ϵ- and higher lactams.

Several lactams are of considerable industrial importance. 2-Pyrrolidinone and 1-methyl-2-pyrrolidinone are made by heating γ-butyrolactone with ammonia and methylamine, respectively. They are useful specialty solvents. Vinylation of 2-pyrrolidinone with acetylene gives 1-vinyl-2-pyrrolidinone, which is polymerized to a substance commonly used in aerosol hair sprays, as in reaction (2).



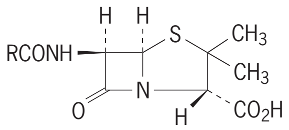
(2)

ϵ-Caprolactam is made commercially by the Beckmann rearrangement of cyclohexanone oxime, and polymerizes when subjected to heat and pressure to form the fibrous polymer nylon-6 [reaction (3)].

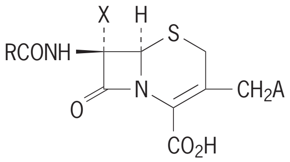


(3)

The β-lactam antibiotics comprise two groups of clinically important therapeutic agents, the penicillins (1) and the cephalosporins (2). In both cases they contain a four-membered or β-lactam ring which has its nitrogen atom and a carbon atom in common with another ring. Such substances are derived commercially from fermentation processes, followed usually by chemical manipulation of the functional groups. *See* AMINO ACIDS; BIOCHEMICAL ENGINEERING; LACTONE; TAUTOMERISM.



(**1**)



(**2**)

Paul E. Fanta

Bibliography. *Kirk-Othmer Encyclopedia of Chemical Technology*, 4th ed., 1994; J. Kroschwitz and M. Winokur, *Chemistry: General, Organic, Biological*, 2d ed., 1990; A. Streitwieser, C. H. Heathcock, and E. M. Kosower, *Introduction to Organic Chemistry*, 4th ed., 1998.

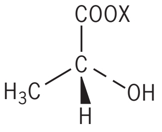
Lactase

An enzyme found in mammals, honeybee larvae, and some plants. It is a β-galactosidase which hydrolyzes lactose to galactose and glucose. In mammals, lactase appears in the intestinal secretion from the intestinal villi, and exerts its effect on lactose in chyme. *See* ENZYME; GLUCOSE.

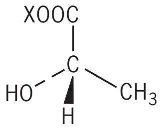
Daniel N. Lapedes

Lactate

A salt or ester of lactic acid ($CH_3CHOHCOOH$). In lactates, the acidic hydrogen of the carboxyl group has been replaced by a metal or an organic radical. Lactates are optically active, with a chiral center at carbon 2. The two optical isomers of a metal (X) lactate have structures (**1**) and (**2**). Commercial fermentation produces either the dextrorotatory (*R*) or the levorotatory (*S*) form, depending on the organism involved. *See* OPTICAL ACTIVITY.



R form (**1**)



S form (**2**)

The *R* form of lactate occurs in blood and muscle as a product of glycolysis. Lack of sufficient oxygen during strenuous exercise causes enzymatic (lactate dehydrogenase) reduction of pyruvic acid to lactate, which causes tiredness, sore muscles, and even muscle cramps. During renewed oxygen supply (rest) the lactate is reoxidized to pyruvic acid and the fragments enter the Krebs (citric acid) cycle. The plasma membranes of muscle and liver are permeable to pyruvates and lactates, permitting the blood to transport them to the liver (Cori cycle). Lactates also increase during fasting and in diabetics. *See* BIOLOGICAL OXIDATION; CARBOHYDRATE METABOLISM; CITRIC ACID CYCLE.

Lactates are found in certain foods (sauerkraut), and may be used for flour conditioning and in food emulsification. Alkali-metal salts act as blood coagulants and are used in calcium therapy, while esters are used as plasticizers and as solvents for lacquers. *See* ESTER; SALT (CHEMISTRY). Elbert H. Hadley

Bibliography. J. R. Holum, *Fundamentals of General, Organic, and Biological Chemistry*, 6th ed., 1997; J. Kroschwitz and M. Winokur, *Chemistry: General, Organic, Biological*, 2d ed., 1990; A. Streitweiser, Jr., C. H. Heathcock, and E. M. Kosower, *An Introduction to Organic Chemistry*, 4th ed., 1998; L. Stryer, *Biochemistry*, 4th ed., 1995.

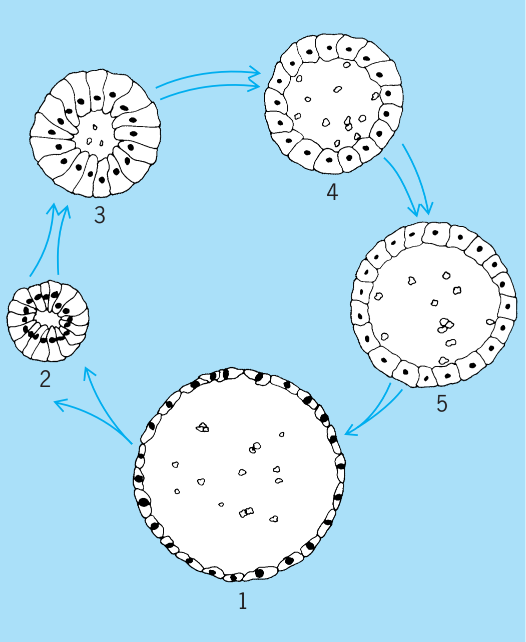
Lactation

The function of the mammary gland providing milk nourishment to the newborn mammal. This process is under the control of the endocrine and nervous systems. It involves transformation of an inactive duct system to a lobuloalveolar glandular structure during pregnancy, cellular production of the components of milk (galactopoiesis), secretion into the ducts, and ejection under the stimulus of milking or suckling.

Galactopoiesis. Formation of the components of milk by the cells composing the lobuloalveolar glandular structure is under the control of a number of hormones. The transformation of the duct structure to a glandular structure during pregnancy initially requires the action of estrogen to induce further duct growth, and of progesterone to induce the development of the lobuloalveolar glandular structure. In addition, somatotropin (STH, growth hormone), corticotropin (ACTH), thyrotropin (TSH), prolactin (PRL) secreted by the anterior lobe of the pituitary (adenohypophysis), and insulin secreted by the pancreas are essential for normal development of the secretory activity of the evolving glandular structure. Cortisone and thyroxine may be substituted for ACTH and TSH. The placenta becomes the source of the estrogen and progesterone and of a mammary-gland hormone which, in humans, is called human placental lactogen (human chorionic somatomammotropin, hCM). Although the gland becomes functional, secretion of milk is minimal until after parturition, when release from the inhibition exerted by the estrogen and progesterone allows secretory activity to rise. Once lactation is initiated after parturition, it is maintained by prolactin with a minimal but necessary contribution from the other pituitary hormones or their target-gland hormones. Milk must be removed by suckling or milking, or the pressure which builds up in the alveoli and ducts will cause retardation of further milk production and involution of the glands.

There is a close correlation between milk secretion and blood levels of prolactin. The initiation of suckling causes a rapid secretion of prolactin and maintenance of the mammary-gland epithelium. The mechanism by which prolactin is thought to effect galactopoiesis involves an interaction of the hormone with a receptor-site protein in the surface of milk secretory cells of the mammary-gland alveoli. The prolactin-receptor site complex is internalized into the cell and, by an undetermined mechanism, enters the nucleus, where it stimulates a genome to transcribe messenger RNA. This nucleic acid then enters the cytoplasm, where the formation it contains is translated into milk protein. The resulting product appears in the lumen of the alveoli as milk (see **illus.**). Between nursings, the blood levels of prolactin are in the range of nonnursing women. Maintenance of the secretory state of the lactating mammary gland requires intermittent stimulation by suckling. Increasing the time between feedings results in less frequent stimulation by prolactin and gradual reduction in milk synthesis and secretion. Secretion of prolactin is held under continuous check by a release-inhibiting hormone produced by the hypothalamus. The suckling stimulus relieves this inhibition and initiates secretion of a releasing hormone.

Lactation makes demands on the maternal regulation of calcium metabolism. Resorption of bone increases in lactating rats and women, and there is a marked increase in the absorption of calcium from the intestine. The elevated need for calcium results in an increased role for parathyroid hormone,



Milk secretory cell activity cycle. (1) The milk-filled alveolus lined with flattened "resting" cells. (2–3) With emptying, as the alveolus contracts and its capillary bed expands, the cells become columnar and resume milk synthesis. (4–5) As milk accumulates in the alveolar lumen, the milk secretory cells regress in height, becoming cuboidal. With further accumulation of milk, the cells reenter their "resting" phase (1). (*After J. M. Nolin and E. M. Bogdanove, Biology of Reproduction, 22:393–416, 1980*)

calcitonin, and vitamin D in the regulation of the absorption and utilization of calcium. In humans a concomitant phenomenon frequently associated with lactation is amenorrhea. Consequently in some societies prolonged nursing is used as a birth control technique.

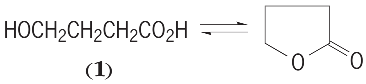
Milk ejection. Discharge of the milk contained in lactating mammary glands is mediated by a reflex. Suckling stimulates receptors in the nipple that induce impulses in sensory nerves leading into the brain. This stimulus causes the hypothalamus to release oxytocin via its nerves terminating in the neurohypophysis (posterior lobe of the pituitary). The oxytocin is transported by the blood to the basketlike arrangement of myoepithelial cells surrounding the mammary alveoli. The contraction of these cells causes milk to be ejected. Farmers refer to this process as “letdown.” Other factors associated with the act of nursing or milking may initiate release of oxytocin. Thus in humans, sexual activity, the approach of the usual nursing time, or preparation of the infant for nursing may become adequate stimuli for release of oxytocin. Analogously, in cows, the sound of buckets, the presence of calves, or washing of the udder may stimulate oxytocin release. Emotional states such as fear or anger which result in secretion of epinephrine may inhibit lactation by inhibiting the action of oxytocin on the myoepithelial basket cells. Nursing mothers should be content, free of stress, and involved primarily with the nursling.

Milk production. The first milk, or colostrum, secreted before parturition, is especially rich in gamma globulin, which contains immune bodies. The digestive tract of the newborn absorbs this protein intact into the blood; this confers a temporary immunity to common diseases. With the regular removal of milk, the amount of milk secreted increases for a period, and then gradually declines until the animal is said to dry up. In dairy animals such as the cow, pregnancy may be induced about the ninetieth day of the lactation period. Lactation and pregnancy with additional growth of the mammary gland occur, so that a second lactation period is initiated in a single year. The total yield to milk increases each year until about the seventh or eighth year, and then declines with old age. *See* ADRENAL GLAND; MAMMARY GLAND; MILK; PITUITARY GLAND; THYROID GLAND. **Harry J. Lipner**

Bibliography. B. L. Larson, *Lactation*, 1985, reprint 1996; T. B. Mepham, *Physiology of Lactation*, 1987; B. S. Worthington et al., *Nutrition in Pregnancy and Lactation*, 4th ed., 1988.

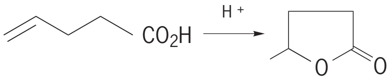
Lactone

A cyclic, intramolecular ester derived from a hydroxy acid. Simple lactones are designated α, β, γ, δ and so forth according to the position of the hydroxyl group, as shown in (**1**). Five- and six-membered lactones are very readily obtained by cyclization of a hydroxy acid or precursor as shown in reactions (1)–(3). For a γ- or δ-hydroxy acid, the equilibrium in reaction (1) usually lies far to the right, favoring the lactone. *See* ESTER.

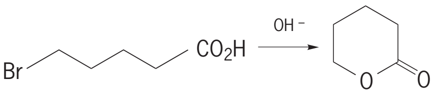


(**1**)

(1)

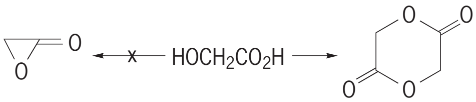


(2)



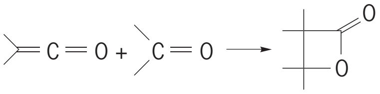
(3)

Lactones with three- and four-membered rings are also known. α-Lactones require special preparative methods since, as shown in reaction (4), heating an α-hydroxy acid leads to the dimeric product, known as a lactide, containing two ester groupings in a six-membered ring (**2**). β-Lactones can be prepared from the halo acid or by cycloaddition of a ketene and carbonyl compound [reaction (5)]. The classical method for seven-membered and larger lactones was oxidation of cyclic ketones or cyclic ethers [reaction (6)], since heating a long-chain hydroxy acid gives a polymeric product.

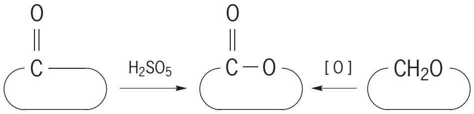


(**2**)

(4)

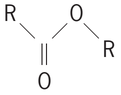


(5)

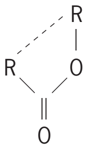


(6)

Lactones undergo the usual reactions of esters such as nucleophilic attack at the carbonyl group and carbanion reactions of the α-methylene group. The strained α- and β-lactones undergo ring opening at the alkyl-oxygen bond. Although the ring in γ- or δ-lactones is not significantly strained, carbonyl attack is much more rapid than that in open-chain esters (**3**) because the conformation of the ester group in the lactone must adopt the higher-energy Z conformation (**4**).

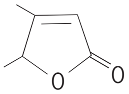


(**3**)

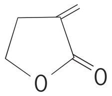


(**4**)

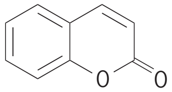
Lactones in various forms are found in numerous naturally occuring compounds. Unsaturated γ-lactone rings (**5** and **6**) are present in many components of essential oil. The δ-lactone moiety is present in coumarin (**7**) and other naturally occuring aromatic compounds. Ascorbic acid [vitamin C; (**8**)] is a carbohydrate lactone.



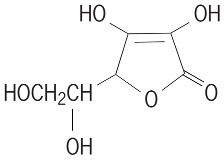
(**5**)



(**6**)



(**7**)



(**8**)

See ASCORBIC ACID; ESSENTIAL OILS.

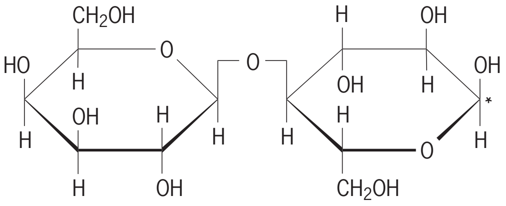
A major development in the chemistry of naturally occuring compounds has been the isolation from microorganisms of a number of macrocylic lactones with rings containing from 12 to more than 30 atoms. These substances include immunosuppressive agents and antibiotics such as erythromycin. The structures of these macrocylic lactones, or macrolides, are extremely complex, with many chiral centers and often several pendant sugar groups attached as glycosides. Although they represent some of the most complex organic compounds known, several of these macrocylic lactones have been synthesized. The most commonly used approach to systhesis is the assembly of subunits to form a hydroxy acid followed by ring closure by means of a thio ester or other activated carboxyl group. *See* ORGANIC CHEMISTRY; ORGANIC SYNTHESIS.

Paul E. Fanta; James A. Moore

Bibliography. J. W. Apsimon (ed.), *Total Synthesis of Natural Products*, vol. 7, 1988; M. Smith and J. March, *March's Advanced Organic Chemistry*, 5th ed., 2000.

Lactose

Milk sugar or 4-*O*-β-D-galactopyranosyl-D-glucose. This reducing disaccharide is obtained as the β-D-anomer (with the formula shown, where the asterisk indicates a reducing group); the melting point (mp) is 202°C (395.6°F) and the optical activity is $[\alpha]_D^{20} + 85.0 \rightarrow +52.6^\circ$. Crystallization at higher temperatures (above 93.5°C or 200.3°F) gives the β-D anomer; mp 252°C (486°F), and $[\alpha]_D^{20} + 35 \rightarrow +55.5^\circ$. Lactose is found in the milk of mammals to the extent of approximately 2–8%. It is usually prepared from whey, which is obtained as a by-product in the manufacture of cheese. Upon concentration of the whey, crystalline lactose is deposited. Lactose is not fermentable by ordinary baker's yeast. In the souring of milk, *Lactobacillus acidophilus* and certain other microorganisms bring about lactic acid fermentation by transforming the lactose into lactic acid, $CH_3CHOHCOOH$. *See* CHEESE; MILK; OPTICAL ACTIVITY.



Chemical evidence shows that the glycosidic linkage involves the carbon atom 1 of D-galactose and carbon 4 of D-glucose. Enzymatic studies indicate that the galactosidic linkage has the β configuration. *See* OLIGOSACCHARIDE.

The mammary glands of lactating animals, and their milk, contain an enzyme, lactose synthetase, capable of transferring the D-galactose unit from uridine diphosphate D-galactose to D-glucose, forming lactose according to the scheme: uridine diphosphate D-galactose + D-glucose → lactose + uridine diphosphate.

The lactose synthetase can be resolved into two protein components, A and B, which individually do not exhibit any catalytic activity. Recombination of these fractions, however, restores full lactose synthetase activity. The B fraction was identified as α-lactalbumin. *See* LACTATION; URIDINE DIPHOSPHOGLUCOSE (UDPG).

William Z. Hassid

Bibliography. W. W. Pigman and D. Horton (eds.), *The Carbohydrates*, 2d ed., vols. 1–2, 1970–1980.

Laevicaudata

An order of fresh-water branchiopod crustaceans formerly included in the order Conchostraca. The body is up to about 7 mm (0.3 in.) in length; females have 12 trunk segments, males have 10. The trunk terminates in a feebly developed telson that lacks claws and is covered ventrally by an opercular lamella. The trunk and its appendages are covered by a bivalved carapace that has a true hinge dorsally and that usually lacks lines of growth. The head is free, articulated with the trunk, and can be swung forward. It is covered by a headshield and has paired, sessile eyes. It bears large antennae that are used for swimming. The mandibles, while of the rolling, crushing type, have narrow masticatory surfaces and stout teeth.

The 12 pairs of trunk limbs of the female are foliaceous. The first, and sometimes the second, of the 10 pairs of the male is modified for clasping the female. Clumps of small drought-resistant eggs are carried by the female and shed at the next molt. They hatch as a late nauplius of distinctive type. The Laevicaudata are inhabitants of temporary water bodies and are almost worldwide in distribution. Although they can swim, they are usually associated with the bottom. *See* BRANCHIOPODA. Geoffrey Fryer

Lagomorpha

An order of mammals composed of two living families: Leporidae (rabbits and hares) and Ochotonidae (pikas). Rabbits and hares have short tails and long ears, whereas pikas lack an external tail and are short-eared. Rabbits and hares occupy a wide variety of habitats and are nearly worldwide in distribution. Pikas are found only in the mountains of western North America, in Siberia, and in the Himalayan region.

Description. Lagomorphs were long considered to be members of the order Rodentia because of their chisel-like incisors, lack of canine teeth, and the presence of a space between the incisors and molariform teeth (known as a diastema). At birth, lagomorphs have three pairs of upper incisors, but the outer on each side is soon lost. In adults, the second pair of incisors is small, is nearly circular, lacks cutting edges, and is set directly behind (not beside) the first functional pair. Each of the anterior upper incisors possesses a longitudinal groove on its anterior face. As in rodents, the incisors grow throughout life, but unlike rodent incisors, they are covered both anteriorly and posteriorly by enamel. In rodents, the enamel is restricted to the anterior face. The distance between the tooth rows of the lower jaws is less than that between the tooth rows of the upper jaws, so only one pair of upper and lower tooth rows is opposable at the same time. The motion of the jaws in mastication is lateral, unlike the rotary jaw movement of rodents. The palate is short and does not extend to the end of the tooth row. It is generally accepted that the resemblances between lagomorphs and rodents are results of convergence rather than of close taxonomic relationship. There is some indication that lagomorphs are more closely related to hooved animals than to rodents. *See* DENTITION; RODENTIA; TOOTH.

The two main lower leg bones, the tibia and fibula, are fused for more than half their length, and the fibula articulates with the calcaneum in the ankle as in artiodactyls. The forelimbs have five digits, while the hindlimbs may have four or five. *See* ARTIODACTYLA.

The diet of lagomorphs is restricted to vegetation. Two types of fecal pellets are expelled: brown pellets and greenish pellets. The reingestion of partially digested vegetation from the greenish pellets is common; this practice is known as coprophagy. It allows the animals to spend relatively little time exposed to predators while in the field actually feeding. They consume green vegetation rapidly and then make optimum use of it in the safety of their brushpile or burrow. This process is somewhat analogous to cud chewing in ruminant animals. These pellets have a high protein content and contain large amounts of B vitamins produced by intestinal bacteria. *See* DIGESTIVE SYSTEM.

Leporidae. This family of 11 living genera and 54 species has a natural range covering most of the world's major landmasses, with the notable exceptions of southern South America, the West Indies, Madagascar, and most islands to the southeast of Asia. Leporids, however, are among the most widely introduced mammals, having been brought by humans to southern South America, Australia, New Zealand, Java, and various oceanic islands. Members of this family have long hindlimbs, oversized feet, furry tails, and long ears. Each limb has five clawed digits, but the first digit is very small. The soles of the feet are covered with hair. A sensory pad, usually hidden by hair-covered folds of skin, is located at the entrance of each nostril, and a Y-shaped naked groove extends from the upper lip to and around the nose. The upper lip is divided. Female leporids are usually larger than males.

The terms "rabbit" and "hare" are often mistakenly interchanged. True rabbits (*Pentalagus, Bunolagus, Pronolagus, Nesolagus, Romerolagus, Brachylagus, Sylvilagus, Oryctolagus, Poelagus*, and *Caprolagus*) prefer habitats with plenty of cover. Young rabbits are born in a well-defined nest. At birth they are altricial; that is, they are blind and naked, have undeveloped ears, and are helpless. Hares (*Lepus*) are bigger than rabbits, have longer hindlegs developed for leaping, and have longer, broader ears. They favor wide-open spaces. Hares, often called jackrabbits in North America, have longer gestation periods than rabbits. The female hare rarely constructs a nest for her young which, at birth, are precocial; that is, they are well developed in that they are furred, have their eyes and ears open, and are able to move about on their own shortly after birth.

Leporids are prolific breeders. Reproductive efficiency is increased by a phenomenon known as induced ovulation. Female leporids normally ovulate only after copulation has occurred, a system that ensures optimum fertilization of ova. *See* REPRODUCTIVE BEHAVIOR; REPRODUCTIVE SYSTEM.

Rabbits and hares are crepuscular, active mainly near dusk and dawn. They are not physically capable of defending themselves against predators, but rely on a keen sense of hearing and smell, the use of burrows, or speed and agility to escape danger. At high speeds, lagomorph limbs function as digitigrade (walking on the digits) appendages, but when moving slowly the entire foot makes contact with the ground (plantigrade).

The Old World, or domestic, rabbit (*Oryctolagus*) has been widely introduced throughout the world. Its spectacular rate of reproduction, especially in areas where no natural predators existed, has proven to be disastrous to other fauna. Nesting areas for petrels and albatrosses on some Pacific islands have been virtually destroyed by the removal of vegetation

Pika (*Ochotona princeps*). (© *California Academy of Sciences*)

by these rabbits and resulting erosion. These rabbits were introduced into Australia by British settlers in 1859 and to New Zealand a few years later. They reproduced until they numbered in the hundreds of millions, causing an ecological disaster in the southern half of Australia. Through their competition and habitat destruction, they have caused the extinction of several native marsupial species and have seriously harmed the sheep-raising industry. In the early 1950s, scientists released myxomatosis, a rabbit-killing virus which proved to be partially successful in controlling the rabbit population but gradually became less effective. In 1995, a calicivirus known as rabbit hemorrhagic disease (RHD) began to be released. By 1997, the wild rabbit population had dropped by 95% in some regions, and native fauna and flora were already staging a comeback. The expectation is that, after the calicivirus kills most of the rabbits, it will remain in the reduced population and act as a long-term regulator of the rabbit population. *See* INFECTIOUS MYXOMATOSIS.

Ochotonidae. Pikas (*Ochotona*), with their stocky bodies and short legs, resemble guinea pigs (see **illustration**). In North America, other common names include cony, rock cony, rock rabbit, mouse hare, whistling hare, little chief hare, and piping hare. While most of the 26 species that occur worldwide inhabit talus, or piles of broken rock, in mountainous regions, several species in Russia, Mongolia, Siberia, and northern China inhabit open plains, deserts, and steppes, and excavate burrows for shelter. Pikas are primarily active during the day, in all seasons. They feed on a variety of vegetable matter. Individuals in most populations spend many hours every day harvesting grasses and herbs and curing them in the sun in exposed locations. These hay piles or haystacks are then stored to be used as food during the winter. As in lagomorphs, coprophagy allows pikas to obtain certain vitamins that would otherwise be lost. Although some species are gregarious and live in large colonies, adult males and females of the two North American species (*O. collaris* and *O. princeps*) are territorial and live separately for most of the year. Most pikas breed twice a year and produce 1–5 young after a gestation of approximately 30 days. Distinctive whistling vocalizations are used during courtship. Rock-dwelling species may live up to 7 years, whereas ground-dwelling species usually live only 1–2 years. Donald W. Linzey

Bibliography. G. A. Feldhamer, B. C. Thompson, and J. A. Chapman, *Wild Mammals of North America: Biology, Management, and Conservation*, 2d ed., Johns Hopkins University Press, 2003; D. Macdonald (ed.), *The Encyclopedia of Mammals*, Andromeda Oxford, 2001; R. M. Nowak, *Walker's Mammals of the World*, 6th ed., Johns Hopkins University Press, 1999; D. E. Wilson and S. Ruff (eds.), *The Smithsonian Book of North American Mammals*, Smithsonian Institution Press, 1999.

Lagrange's equations

Equations of motion of a mechanical system for which a classical (nonquantum-mechanical) description is suitable, and which relate the kinetic energy of the system to the generalized coordinates, the generalized forces, and the time. If the configuration of the system is specified by giving the values of f independent quantities $q_1, \ldots, q_f$, there are f such equations of motion. In their usual form, these equations are equivalent to Newton's second law of motion and are differential equations of the second order for the q's as functions of the time t.

Derivation. Let the system consist of N particles. The masses of the particles are m_ρ and their cartesian coordinates are $x_{\rho i}$ ($\rho = 1, 2, \ldots, N$; $i = 1, 2, 3$). These cartesian coordinates are expressible as functions of f ($\leqq 3N$) generalized coordinates $q_1, \ldots, q_f$ between which there are no constraints. The time does not appear here if the constraints are fixed (as will be assumed), as in Eq. (1). Then, denoting time

$$x_{\rho i} = f_{\rho i}(q_1, \ldots, q_f) \tag{1}$$

differentiation by a dot, Eq. (2) holds. The cartesian

$$\dot{x}_{\rho i} = \sum_{j=1}^{f} \frac{\partial x_{\rho i}}{\partial q_j} \dot{q}_j \tag{2}$$

velocity components are linear functions of the generalized velocities q_j.

Let δq_j represent a small displacement of the system. It is automatically consistent with the constraints. In this displacement, the forces of constraint do no work; their only action is to prevent motion contrary to the constraint, so they have no components in the direction of a displacement consistent with the constraints. If W is the work done during the displacement, it is done entirely by the externally applied forces, as in Eq. (3).

$$W = \sum_{\rho,i} F_{\rho i}\delta x_{\rho i} = \sum_{\rho,i} m_\rho \ddot{x}_{\rho i}\delta x_{\rho i} \tag{3}$$

Each term in the sum of Eq. (3) may receive contributions from forces of constraint, but they cancel when the summation is made. From Eq. (1), Eqs. (4) and (5) are obtained.

$$\delta x_{\rho i} = \sum_{j} \frac{\partial x_{\rho i}}{\partial q_j} \delta q_j \tag{4}$$

$$\frac{\partial \dot{x}_{\rho i}}{\partial \dot{q}_j} = \frac{\partial x_{\rho j}}{\partial q_j} \tag{5}$$

It is readily verified that Eq. (6) is valid.

$$W = \sum_j \left[\frac{d}{dt} \frac{\partial}{\partial \dot{q}_j} \left(\frac{1}{2} m_\rho \dot{x}_{\rho i}{}^2 \right) - \frac{\partial}{\partial q_j} \left(\frac{1}{2} m_\rho \dot{x}_{\rho i}{}^2 \right) \right] \delta q_j \tag{6}$$

The quantity in the brackets is thus, for each j, the noncartesian analog of the carterian $m_\rho \ddot{x}_{\rho i}$.

The only quantity which appears differentiated in Eq. (6) is the total kinetic energy T of the system. This is easily calculated in generalized coordinates because the connection between the cartesian velocities and the generalized velocities is linear and homogeneous. Usually, the kinetic energy can be written by inspection without using Eq. (2) explicitly. Thus Eq. (7) holds.

$$W = \sum_j \left(\frac{d}{dt} \frac{\partial T}{\partial \dot{q}_j} - \frac{\partial T}{\partial q_j} \right) \delta q_j \tag{7}$$

Transforming the right-hand side of Newton's equation is simpler, as shown in Eqs. (8) and (9),

$$W = \sum_{\rho, i} F_{\rho i} \delta x_{\rho i} = \sum_j Q_i \delta q_j \tag{8}$$

$$Q_j = \sum_{\rho, i} F_{\rho i} \frac{\partial x_{\rho i}}{\partial q_j} \tag{9}$$

where Eq. (9) is the jth component of the generalized force. By the preceding argument, Q_j depends only on the externally applied forces, the forces of constraint necessarily canceling in the summation.

The displacement δq_j was entirely arbitrary. Thus, it follows from equating expressions (7) and (8) for W that Eq. (10) is valid. Equation (10) sums up

$$\frac{d}{dt} \frac{\partial T}{\partial \dot{q}_j} - \frac{\partial T}{\partial q_j} = Q_j \tag{10}$$

Lagrange's equations of motion. They are valid also when moving constraints are present. *See* CONSTRAINT.

Examples. Use of this form of Lagrange's equations is shown in the two following examples.

Particle in central force field. Here, the force acting on a particle acts always through a fixed point. Choose this point as origin of a spherical coordinate system with coordinates r, θ, ϕ (**Fig. 1**). Only radial displacements involve work, so that only Q_r differs from zero, or Eq. (11) holds.

$$Q_\theta = Q_\phi = 0 \tag{11}$$

The kinetic energy is given by inspection of Fig. 1 and is expressed by Eq. (12).

$$T = \frac{m}{2} (\dot{r}^2 + r^2 \dot{\theta}^2 + r^2 \sin^2 \theta \, \dot{\phi}^2) \tag{12}$$

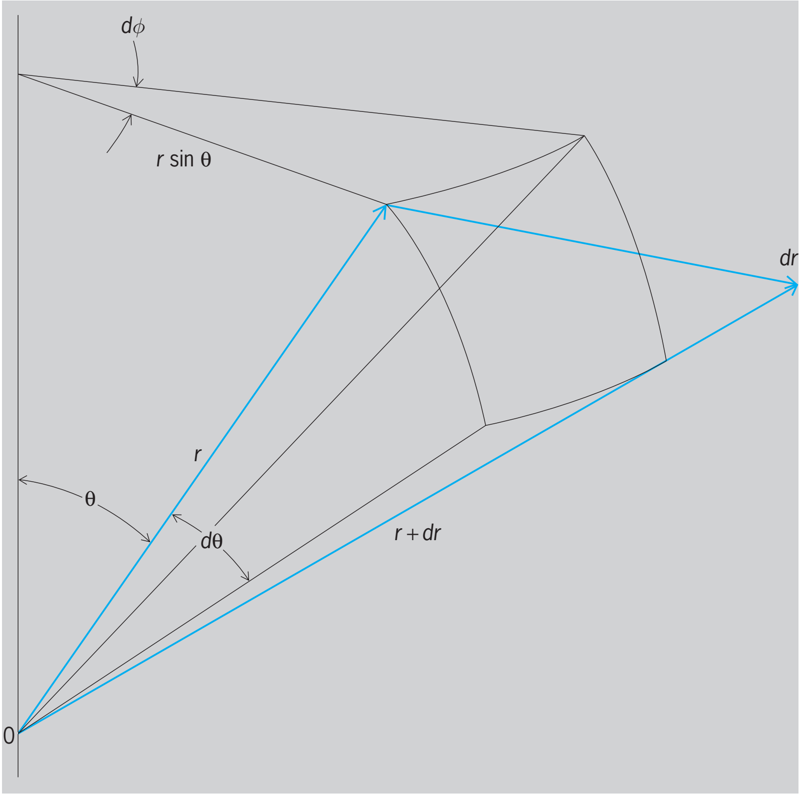


Fig. 1. The vector *dr* is decomposed into three orthogonal vectors of magnitudes *r* sin θ $d\phi$, $rd\theta$, and *dr*, respectively, *r*, θ, and ϕ being spherical coordinates of the terminus of the vector *r*.

Lagrange's equations are Eqs. (13). These may

$$\frac{d}{dt} \frac{\partial T}{\partial \dot{r}} - \frac{\partial T}{\partial r} = m(r - r\dot{\theta}^2 - r \sin^2 \theta \, \dot{\phi}^2) = Q_r$$

$$\frac{d}{dt} \frac{\partial T}{\partial \dot{\theta}} - \frac{\partial T}{\partial \theta} = m \left[\frac{d}{dt} (r^2 \dot{\theta}) - r^2 \sin \theta \cos \theta \, \dot{\phi}^2 \right] = 0 \tag{13}$$

be compared with the cartesian equations (14), in

$$\begin{aligned} m\ddot{x} &= \frac{x}{r} F(x, y, z) \\ m\ddot{y} &= \frac{y}{r} F(x, y, z) \\ m\ddot{z} &= \frac{z}{r} F(x, y, z) \end{aligned} \tag{14}$$

which the force function F appears in all three of the equations of motion, while two of the three Lagrange equations are independent of the detailed nature of the force.

Two particles, fixed separation. In this system, there is one constraint, the particles being a constant distance d apart, and so there are five degrees of freedom instead of six. Choose as generalized coordinates the cartesian coordinates X, Y, Z of the center of mass and the polar angles θ, ϕ of the line joining the two particles, as in **Fig. 2**. Then because the kinetic energy is the sum of the kinetic energy of the center of mass and the kinetic energy relative to

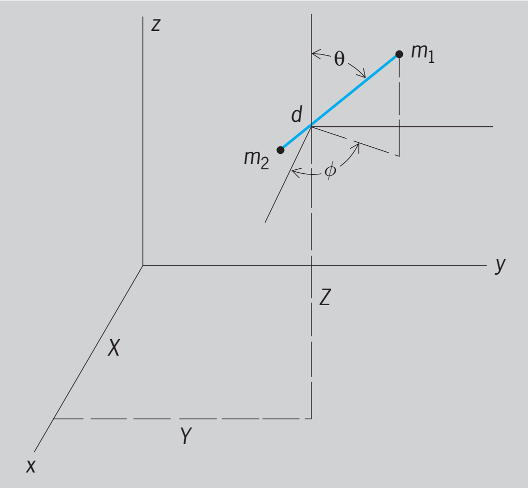


Fig. 2. ***X*, *Y*, and *Z* are cartesian coordinates of the center of mass of dumbbell m_1m_2; θ and ϕ are the polar angles of the dumbbell axis in a set of axes parallel to fixed set of axes and with origin at center of mass.**

the center of mass, Eq. (15) holds. The equations of

$$T = \frac{m_1 + m_2}{2}(\dot{X}^2 + \dot{Y}^2 + \dot{Z}^2) + \frac{1}{2}\left(\frac{m_1 m_2}{m_1 + m_2}\right) d^2(\dot{\theta}^2 + \sin^2\theta\ \dot{\phi}^2) \quad (15)$$

motion may be written down immediately, the generalized forces being evaluated from Eq. (9). There is one equation for each degree of freedom, and the constraint is automatically satisfied.

Conservative systems. In many problems, the forces Q_j are derivable from a potential. Then Eq. (16), and thus Eq. (17), is valid. When this is

$$W = \sum_{j=1}^{f} Q_j dq_j = -dV \quad (16)$$

$$Q_j = -\frac{\partial V}{\partial q_j} \qquad \frac{\partial V}{\partial \dot{q}_j} = 0 \quad (17)$$

so, it is convenient to define a function, called the Lagrangian, by Eq. (18). Then the equations of

$$L(q, \dot{q}t) = T(q, \dot{q}, t) - V(q, t) \quad (18)$$

motion become simply Eq. (19), which is the

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{q}_j} - \frac{\partial L}{\partial q_j} = 0 \quad (19)$$

most commonly encountered form of Lagrange's equations.

Example of use of L. A simple vibration problem illustrates the use of generalized coordinates and shows the ease of application of Lagrange's equations when T and V are relatively easy to obtain. Consider the one-dimensional system of two equal masses m connected by a spring of restoring constant k. Let the displacements of the masses from their equilibrium positions be x_1 and x_2, respectively.

Now introduce new coordinates, $q_1 = x_2 - x_1$ and $q_2 = x_2 + x_1$. Then Eqs. (20) hold.

$$\begin{aligned} V &= \frac{kq_1{}^2}{2} \\ T &= \frac{m(\dot{q}_1{}^2 + \dot{q}_2{}^2)}{4} \end{aligned} \quad (20)$$

Using $L = T - V$, Lagrange's equations for the two variables are written as Eqs. (21).

$$\begin{aligned} {}^1/_2 m\dot{q}_1 - kq_1 &= 0 \\ {}^1/_2 m\dot{q}_2 = 0 \text{ or } \dot{q}_2 &= 0 \end{aligned} \quad (21)$$

The solution to the first of these differential equations is $q_1 = A \sin\ 2k/m\ t$ (setting $q_1 = 0$ at $t = 0$), which is a simple harmonic vibration of the two masses. The second equation merely states that the acceleration of the center of mass of the system is zero.

Nonconservative systems. For (19) is also correct even if the system is not conservative, provided one can write Eq. (22), as in the case of a charged particle in an electromagnetic field. Here V is defined in Eq. (23), where ϕ and $\mathbf{A}$ are the scalar and vector

$$Q_j = \frac{d}{dt}\frac{\partial V}{\partial \dot{q}_j} - \frac{\partial V}{\partial q_j} \quad (22)$$

$$V = e\left(\phi - \frac{\mathbf{v}\cdot\mathbf{A}}{c}\right) \quad (23)$$

potentials of the field, respectively, e is the charge of the particle, $\mathbf{v}$ is the particle's velocity, and c is the velocity of light.

Cyclic coordinate. If L does not depend explicitly on a particular coordinate, say q_k, then Eqs. (24) hold.

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{q}_k} = 0 \qquad \frac{\partial L}{\partial \dot{q}_k} = \text{constant} \quad (24)$$

Here q_k is an ignorable or cyclic coordinate, and $\partial L/\partial q_k$ is an integral of motion. In the example of the central force field given before, ϕ may be ignored.

Total energy. By use of Eq. (19), Eq. (25) is obtained.

$$\frac{d}{dt}\left(\sum_{j=1}^{f}\frac{\partial L}{\partial \dot{q}_j}\dot{q}_j - L\right) = -\frac{\partial L}{\partial t} \quad (25)$$

Hence, if L does not depend explicitly on the time, expression (26) is an integral of motion. Usually, it

$$\sum_{j=1}^{f}\left(\frac{\partial L}{\partial q_j}q_j - L\right) \quad (26)$$

is the total energy of the system. *See* HAMILTON'S EQUATIONS OF MOTION.

Conjugate momentum. The quantity in Eq. (27) is de-

$$p_k = \frac{\partial L}{\partial \dot{q}_k} \quad (27)$$

fined to be the momentum conjugate to the coordinate q_k. (If a Lagrangian does not exist because of the presence of dissipative forces, the conjugate momentum is sometimes defined as $\partial T/\partial \dot{q}_k$.) This momentum is not necessarily a linear momentum as defined in Newtonian mechanics, since its character depends both on the system and on the nature of the coordinate q_k. If q_k is an ignorable coordinate, its conjugate momentum is a constant of motion.

Kinetic momentum. For a charged particle in an electromagnetic field, the Lagrangian may be written as Eq. (28), and the momentum conjugate to the cartesian coordinate x is given by Eq. (29). The quantity

$$L = {}^1/_2 m\mathbf{v}^2 - e\phi + \frac{e}{c}\mathbf{v}\cdot\mathbf{A} \qquad (28)$$

$$p_x = \frac{\partial L}{\partial \dot{x}} = m\dot{x} + \frac{e}{c}A_x \qquad (29)$$

$m\dot{x}$ is called the kinetic momentum. The kinetic momentum is related to p_x much as the kinetic energy is related to the total energy. If Φ and A are independent of x, it is the momentum p_x and not the kinetic momentum $m\dot{x}$ which is a constant of motion.

Relativistic systems. The equations of motion of a relativistic particle may be written in Lagrangian form. The simplest way to do this is to replace the kinetic energy T by another function τ of the mass and velocity so as to get the desired form, namely, Eq. (30). Here m_0 is the rest mass of the particle.

$$\frac{d}{dt}\frac{\partial \tau}{\partial \dot{x}_k} - \frac{\partial V}{\partial x_k} = \frac{d}{dt}\left(\frac{m_0\dot{x}_k}{\sqrt{1 - v^2/c^2}}\right) - \frac{\partial V}{\partial x_k} = 0 \qquad (30)$$

This is accomplished by setting τ as in Eq. (31),

$$\tau = \left(1 - \sqrt{1 - v^2/c^2}\right) m_0 c^2 \qquad (31)$$

which reduces to T in the limit $\mathbf{v}/c \to 0$. The equations of motion in this form are still valid in only one reference frame because the time and the coordinates are treated on different bases. *See* RELATIVISTIC MECHANICS; RELATIVITY. Philip M. Stehle

Bibliography. R. Baierlein, *Newtonian Dynamics*, 1983; V. D. Barger and M. G. Olsson, *Classical Mechanics: A Modern Perspective*, 2d ed., 1995; H. C. Corben and P. Stehle, *Classical Mechanics*, 2d ed., 1960, reprint 1994; H. Goldstein, C. P. Poale, and J. L. Safko, *Classical Mechanics*, 3d ed., 2002; L. D. Landau and E. M. Lifshitz, *Mechanics*, 3d ed., 1976; K. R. Symon, *Mechanics*, 3d ed., 1971.

Lagrangian function

A function of the generalized coordinates and velocities of a dynamical system from which the equations of motion in Lagrange's form can be derived. The lagrangian function is denoted by $L(q_1, \ldots, q_f, \dot{q}_1, \ldots, \dot{q}_f, t)$. *See* LAGRANGE'S EQUATIONS.

For systems in which the forcesare derivable from a potential energy V, if the kinetic energy is T, Eq. (1) holds.

$$L = T - V \qquad (1)$$

If the system is continuous rather than discrete, the lagrangian function L is the integral of a lagrangian density $\mathscr{L}$, as in Eq. (2), where $\eta(x_1, x_2, x_3)$ describes the displacement of the medium at the point (x_1, x_2, x_3). The equations of motion are in this case written as Eq. (3).

$$L = \int \mathscr{L}(\eta, \operatorname{grad}\eta, x_1, x_2, x_3, t)\, dx_1\, dx_2\, dx_3 \qquad (2)$$

$$\frac{\partial}{\partial t}\frac{\partial \mathscr{L}}{\partial(\partial\eta/\partial t)} + \sum_{j=1}^{3}\frac{\partial}{\partial x_j}\frac{\partial \mathscr{L}}{\partial(\partial\eta/\partial x_j)} - \frac{\partial \mathscr{L}}{\partial \eta} = 0 \qquad (3)$$

This formulation of Lagrange's equations applies to the motion of a gas containing sound waves, to a vibrating jelly, or to any medium where discrete masses are replaced by a continuum. Philip M. Stehle

Lake

An inland body of standing water occupying a depression on the Earth's surface. The surface of a lake is exposed to the atmosphere; in addition, subterranean lakes occur. The term lake includes expanded parts of rivers, reservoirs behind dams, and lake basins intermittently or previously covered with water. Small lakes are usually called ponds, tarns (in mountains), and less frequently pools or meres. Lakes are variable in size, with most lakes being less than 65 ft (20 m) in depth. Collectively, lakes cover approximately 9.6×10^5 mi^2 (2.5×10^6 km^2) of the Earth's surface (see **table**). The largest lake basins together hold about 40% of the Earth's freshwater supplies; about half this volume is contained in a single lake basin, Lake Baikal. *See* HYDROSPHERE.

Most lake basins are elliptic sinusoids in cross section, with mean depths about half their maximum depths. Only a few are considered to be extremely deep, with mean depths that exceed 1300 ft (400 m). Although these generally occur in glaciated mountain regions, the Great Lakes of the United States are included in this category. Lakes Baikal and Tanganyika have maximum depths in excess of 3300 ft (1000 m) and mean water depths of 1650 ft (500 m).

Lakes generally cover relatively small areas of the Earth's surface. The largest of the lakes in terms of area is the Caspian Sea, a salt lake, and the largest fresh-water lake is Lake Superior in North America. Together, the five Great Lakes of North America represent the largest continuous mass of fresh water on the Earth's surface, and account for about 95% of the fresh surface water resources in North America.

Formation. Lake-filled depressions form by a great variety of natural processes, including tectonic activity, volcanic activity, landsliding, glacial activity, bedrock solution, river activity, wind activity, shoreline processes, and biological activity.

Dimensions of some major lakes[*]

Lake	Area, mi^2[*]	Volume (approx.), 1000 acre-ft	Shoreline, mi	Depth	
				Av. ft	Max. ft
Caspian Sea	168,500	71,300	3,730	675	3,080
Superior	32,200	9,700	1,860	475	1,000
Victoria	26,200	2,180	2,130		
Aral Sea	26,233[†]	775	—		
Huron	23,010	3,720	1,680		
Michigan	22,400	4,660	—		870
Baikal	13,300[†]	18,700	—	2,300	5,000
Tanganyika	12,700	8,100	—		4,700
Great Bear	11,490[†]	—	1,300		
Great Slave	11,170[†]	—	1,365		
Nyasa	11,000	6,800		900	2,310
Erie	9,940	436			
Winnipeg	9,390[†]		1,180		
Ontario	7,540	1,390			
Balkash	7,115				
Ladoga	7,000	745			
Chad	6,500[†]				
Maracaibo	4,000[†]				
Eyre	3,700[†]				
Onega	3,764	264			
Rudolf	3,475[†]				
Nicaragua	3,089	87			
Athabaska	3,085				
Titicaca	3,200	575			
Reindeer	2,445				

[*] $1\ \text{mi}^2 = 2.6\ \text{km}^2$. $1000\ \text{acre-ft} = 1.2 \times 10^6\ \text{m}^3$. 1 mi = 1.6 km. 1 ft = 0.3 m.

[†] Area fluctuates.

Lake basins develop in areas where the Earth's surface is affected by internal geological processes. These processes result in the vertical displacement of the crust and the subsequent development of lakes in grabens or broadly warped depressions. Examples of such tectonic lakes are Lake Baikal and Lake Tanganyika. Some lake basins are tectonically uplifted above sea level, thus isolating the sea floor. Lakes resulting from such uplift include the Caspian and Aral seas. *See* GRABEN; PLATE TECTONICS.

Volcanic lakes form in a variety of ways. Explosive volcanic activity resulting in the formation of either a maar or a caldera is frequently followed by lake formation. In the case of maars, the lake is formed either by lava coming into contact with ground water or from volcanic degassing. In the case of calderas, the lake forms as a result of the seepage of ground water into the subsided crater of the volcano following complete or partial emptying of the magma chamber. Some volcanic lakes form in essentially tectonic depressions formed as a result of accompanying deformation of the ground surface during eruptions. The third principal means by which volcanic lakes form is by modification of drainage patterns due to molten lava blocking stream flow and impounding river water behind the flows. *See* CALDERA; LAVA; MAGMA; VOLCANO.

Temporary or permanent lakes frequently develop in association with landslide occurrences. Landslide events commonly block streams, causing water to accumulate behind the debris. These lakes are commonly ephemeral, surviving only a short period of time before the stream adjusts to the sudden addition of a mass of sediment and erodes through the debris. In other cases, the lake may remain for some time. Substantially smaller and extremely ephemeral lakes may also develop in depressions in the headwall region of the landslide. These lakes commonly drain after a short period of time. *See* LANDSLIDE.

Perhaps the most common cause of lakes is glacial and periglacial activity which occurs widely in the high-latitude areas of the Northern Hemisphere as well as the high altitudes of both hemispheres. Lakes are associated with both the erosional and depositional processes of glaciation. Lakes fill depressions scoured by alpine glaciers in their source areas to form cirque lakes, or tarns. Lakes also commonly fill depressions scoured by alpine glaciers at progressively lower elevations to form a chain of paternoster lakes. Continental glaciers scour large areas of bedrock to form depressions such as those filled by the Great Lakes of North America. Glacial lakes are commonly impounded behind morainal material left both in the terminal position of glaciers and at temporary stillstand positions of a retreating glacier. Many smaller lakes, called kettle lakes, occupy depressions resulting from the melting of blocks of glacial ice in morainal debris. In addition, small lakes may result from the impoundment of water behind lateral moraine features. Thaw lakes commonly form in permafrost-dominated areas, where ground water seeps into and melts permafrost. *See* GLACIATED TERRAIN; MORAINE; PERMAFROST.

A wide variety of other, less catastrophic processes operate to create lakes. These include lakes that form as a result of ground-water and surface-water filling of solution features in limestone terranes, including dolines and poljes, which are features of karst topography. The erosional and depositional processes of rivers result in a diversity of lake forms along river channels. These include lakes at the base of waterfalls, plunge pools, which may be either active or abandoned. They also include oxbow lakes, which occupy abandoned meander loops of river channels. Coastal-zone lakes commonly form where a coastline embayment becomes blocked by sand transported by longshore drift. In addition, lakes may become entrapped behind coastal dune systems. *See* KARST TOPOGRAPHY; RIVER.

Eolian, or wind-related, processes also create depressions that may be permanently or intermittently filled with water. Interdune corridors may become areas of impoundment. Deflation of desert and semidesert surfaces creates depressions, which often become filled with water to form playas or salinas. *See* EOLIAN LANDFORMS; PLAYA; SALINE EVAPORITES.

Lakes are also created by biological activity. Plant growth due to nutrient enrichment associated with pollution as well as siltation in rivers often leads to lake formation along stream channels. Similarly, beaver activity along streams blocks free flow, and lakes are created. Human activity also leads to the creation of lakes behind a variety of dam structures. *See* EUTROPHICATION; FRESHWATER ECOSYSTEM.

Water characteristics. Lakes are typically subdivided into saline and nonsaline types, with the former including a spectrum of salt contents. Salinity, especially in smaller lakes, is also often variable, fluctuating with seasonal climatic patterns. Typically the salinity of lakes reflects the climatic environment in which they occur. Lakes in arid and semiarid environments, with high potential evaporation, are typically more saline than lakes in humid environments, with low potential evaporations. Examples are the Dead Sea, which has a salinity approximately seven times that of seawater, and the Great Salt Lake in the Western United States, which has a salinity approximately four times that of seawater. Other salt lakes have salinities that are less than that of seawater but are clearly salty to the taste. Freshwater lakes also vary considerably in their salt content, depending on their geologic environment.

Thermal stratification of lakes is a major physical characteristic, reflecting the prevailing climate, lake morphometry, and chemistry; it has profound influences on other physical processes, such as freezing behavior, as well as on the biology of lakes. Most of the world's lakes that are located in temperate latitudes are dimictic; that is, they circulate freely twice a year, with thermal stratification in the summer and inverse stratification in the winter. In the summer the waters of such lakes are warmed from the surface to depth more rapidly than lake-water mixing occurs. In the winter the lake waters cool as heat loss exceeds heat input and inverse stratification occurs. Between these two periods of relative thermal stability, most lakes of moderate depth (>33 ft or 10 m) mix their waters and undergo fall and spring turnover. *See* MEROMICTIC LAKE.

Environmental considerations. Lakes created behind constructions are common features of the Earth's landscape. Such lakes include Lake Mead and Lake Powell on the Colorado River, Lake Roosevelt on the Columbia River, Lake Nasser on the Nile, and Lake Tsimlyanskaya on the Don, which flows into the Black Sea. While major dam construction has largely ceased in the western world, major dam building continues in the developing countries such as Losoto. These lakes, as well as many natural lakes, are significant for the storage and supply of water for irrigation and domestic use, control of floods, generation of electric power, use as commercial waterways, and recreation. However, these diverse uses frequently result in adverse impacts on water quality. Common environmental problems include increases in salinity, siltation of the reservoir, and excessive algal production due to the spillage and inflow of a variety of nutrients. River impoundment also has considerable adverse impact downstream from the dam with respect to flow regimes as well as changes in water temperature. These changes have profound adverse effects on both biota and plant growth. *See* DAM; STREAM TRANSPORT AND DEPOSITION.

While the environmental health of many of the larger lakes has improved considerably with increased global environmental awareness, many are still substantially polluted. For example, there are in excess of 100 toxic substances that can be identified in the waters and sediments of the Great Lakes of North America. While these lakes continue to support a great diversity of life forms, the waters and sediments are contaminated with varying concentrations of chemical elements associated with waste from paper and woodpulp industry as well as from the treatment of sewage. Polychlorinated biphenyls, leaked from electrical capacitors and transformers, continue to be released from sediments on the lake floors and constitute major concerns for human health. Fish production continues to decline in the Great Lakes. *See* WATER POLLUTION. John C. Dixon

Bibliography. A. Lerman (ed.), *Lakes: Chemistry, Geology and Physics*, 2d ed., 1995; J. McManus and R. W. Duck (eds.), *Geomorphology and Sedimentology of Lakes and Reservoirs*, 1993; M. R. Rosen (ed.), *Paleoclimate and Basin Evolution of Playa Systems*, Geol. Soc. Amer. Spec. Pap. 289, 1994; R. G. Wetzel, *Limnology*, 3d ed., 2001.

Lamiales

An order of flowering plants (Magnoliophyta, or angiosperms), in the subclass Asteridae (Eudicotyledons). The order consists of some 22 families with approximately 1100 genera and over 21,000 species. Seven families (Acanthaceae, Gesneriaceae, Lamiaceae, Orobanchaceae, Plantaginaceae, Scrophulariaceae, Verbenaceae) have more than 1000 species. Orobanchaceae, Plantaginaceae, and Scrophulariaceae represent monophyletic groups segregated from the polyphyletic former

Scrophulariaceae *sensu lato*. Members of the Lamiales are distributed worldwide with some families being predominantly tropical (such as Acanthaceae, Bignoniaceae, Gesneriaceae, Verbenaceae) and some being predominantly temperate (such as Oleaceae, Orobanchaceae, Plantaginaceae, Scrophulariaceae). Lamiaceae, the largest family, is well represented in both tropical and temperate floras.

A basal split in the evolutionary relationships within the order sets off the family Oleaceae, with tetramerous flowers, which are radially symmetric, from the rest of the Lamiales. Pentamerous flowers with fewer stamens (two or four) than petals and bilateral symmetry predominate among the other families. Flowers of all families are characterized by petals fused into a tubular corolla, a single whorl of stamens (typically four), and ovaries consisting of two fused carpels. However, the number of ovules per ovary and, therefore, the number of seeds per fruit varies from one to many and helps form the basis for many family distinctions. Iridoid compounds are found in most members of the Lamiales and are thought to function in deterring herbivory by insects. However, iridoids are not universal in the order and, for example, the large subfamily Nepetoideae in family Lamiaceae lacks iridoids but exhibits a diverse array of volatile oils that may serve a similar function.

The volatile oils in the Lamiaceae have made many species valuable as scents and herbs, including lavender (*Lavandula*), mint (*Mentha*), sage (*Salvia*), catnip (*Nepeta*), basil (*Ocimum*), and oregano (*Oreganum*). Additional economically important plants in the Lamiales include those grown for food: sesame (*Sesamum*, Pedaliaceae) and olives (*Olea*, Oleaceae); wood: teak (*Tectona*, Verbenaceae) and ash (*Fraxinus*, Oleaceae); houseplants: African violets and relatives (Gesneriaceae); and many ornamental plants from a wide variety of families. The order also contains some plants with specialized habits, including parasitic plants (Orobanchaceae), insectivorous plants (Byblidaceae, Lentibulariaceae), mangroves (Avicenniaceae), and aquatic plants, such as *Callitriche* and *Hippuris* (Plantaginaceae). *See* ASTERIDAE; MAGNOLIOPHYTA; MAGNOLIOPSIDA; PLANT KINGDOM.

Richard Olmstead

Bibliography. Angiosperm Phylogeny Group, An ordinal classification of the families of flowering plants, *Ann. Missouri Bot. Gard.*, 85:531–553, 1998; R. M. Harley and T. Reynolds (eds.), *Advances in Labiate Science*, Royal Botanic Gardens Kew, Richmond, 1992; S. J. Wagstaff and R. G. Olmstead, Phylogeny of the Labiatae and Verbenaceae inferred from *rbc*L sequences, *Syst. Bot.*, 22:165–179, 1997.

Laminar flow

A smooth, streamline type of viscous fluid motion characteristic of flow at low-to-moderate deformation rates. The name derives from the fluid's moving in orderly layers or laminae without the formation of small eddies or irregular fluctuations.

The chief criterion for laminar flow is a relatively small value—less than a few thousand—for the Reynolds number, Re $= \rho VL/\mu$, where ρ is fluid density, V is flow velocity, L is body size, and μ is fluid viscosity. Thus laminar flow may be achieved in many ways: low-density flows as in rarefied gases; low-velocity or "creeping" motions; small-size bodies such as microorganisms swimming in the ocean; or high-viscosity fluids such as lubricating oils. At higher values of the Reynolds number, the flow becomes disorderly or turbulent, with many small eddies, random fluctuations, and streamlines intertwining like spaghetti. *See* CREEPING FLOW; REYNOLDS NUMBER; TURBULENT FLOW; VISCOSITY.

Nearly all of the many known exact solutions of the equations of motion of a viscous fluid are for the case of laminar flow. These mathematically accurate descriptions can be used to give insight into the more complex turbulent and transitional flow patterns for which no exact analyses are known. *See* NAVIER-STOKES EQUATION.

The theory of viscous lubricating fluids in bearings is a highly developed area of laminar flow analysis. Even large Reynolds number flows, such as aircraft in flight, have regions of laminar flow near their leading edges, so that laminar flow analysis can be useful in a variety of practical and scientifically relevant flows. *See* ANTIFRICTION BEARING; BOUNDARY-LAYER FLOW; LUBRICANT.

Frank M. White

Bibliography. D. F. Rogers, *Laminar Flow Analysis*, 1992; H. Schlichting, *Boundary Layer Theory*, 8th ed., 2000; F. S. Sherman, *Viscous Flow*, 1990; F. M. White, *Viscous Fluid Flow*, 2d ed., 1991.

Laminariales

An order of large brown algae (Phaeophyceae) commonly called kelps. Definitive features include a life history in which microscopic, filamentous, dioecious gametophytes alternate with a massive, parenchymatous sporophyte; growth of the sporophyte effected by meristems; and production of unilocular zoosporangia in extensive sori on blades of the sporophyte. *See* PHAEOPHYCEAE.

General structure. A mature sporophyte (**Fig. 1***a*) consists typically of a holdfast, stipe, and one or more blades. The holdfast usually comprises a cluster of rootlike structures (haptera); in a few species it is discoid or conical. The stipe, which varies in length from a few millimeters in *Hedophyllum* to more than 100 ft (30 m) in *Macrocystis* (giant kelp), may be branched or unbranched and may bear one or more blades. In some genera, fertile blades are distinct from vegetative blades. In the transition region between the stipe and the blade there is a meristem, which adds material to the upper end of the stipe and the lower end of the blade. The girth of the stipe is increased by activity of a superficial meristem (meristoderm). In some genera the sporophyte is annual; in others the stipe is perennial and develops annual growth rings while the blade is regenerated annually

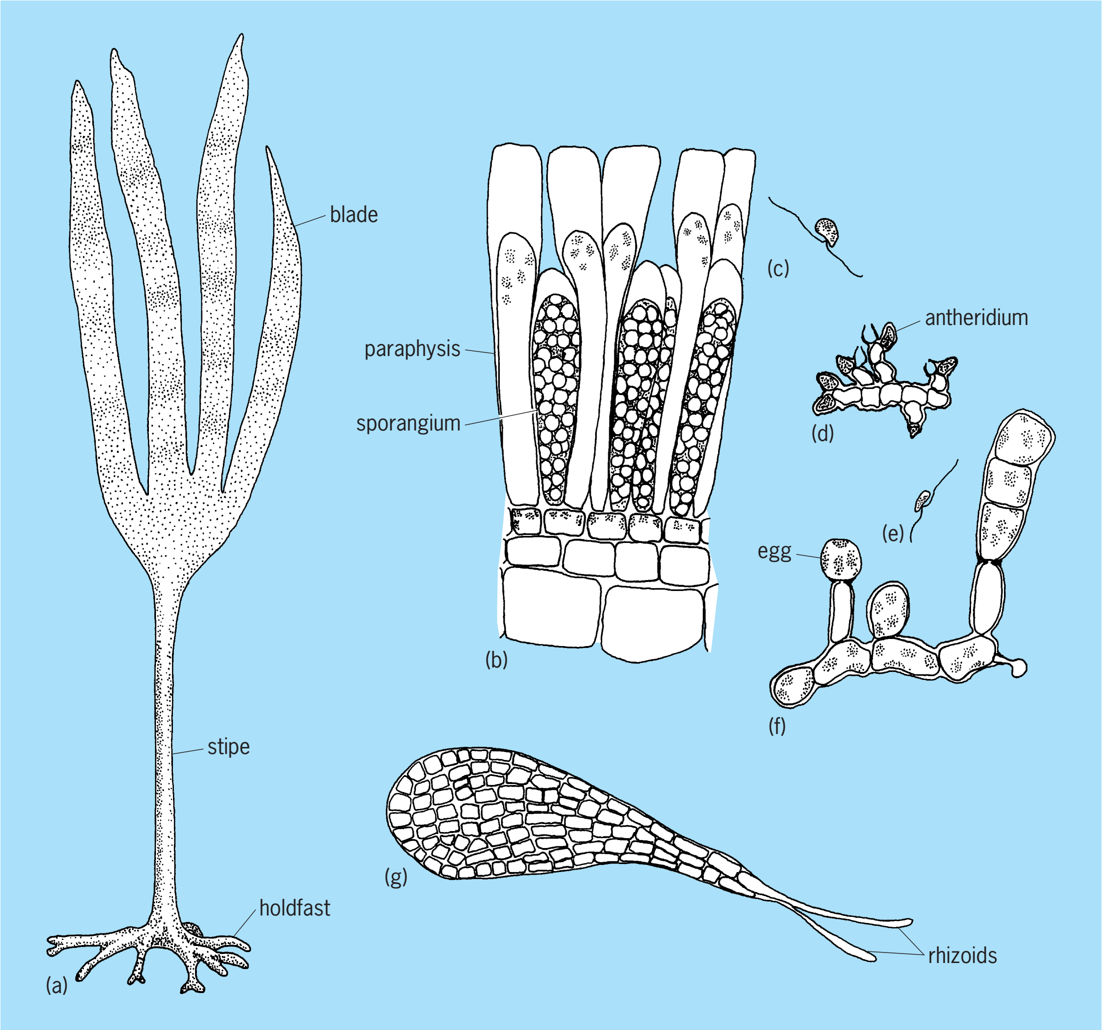


Fig. 1. ***Laminaria*: (*a*) sporophyte; (*b*) cross section of blade, showing sorus of unilocular sporangia interspersed with sterile hairs (paraphyses); (*c*) zoospore; (*d*) male gametophyte with antheridia; (*e*) antherozoid (sperm); (*f*) female gametophyte with egg (left) and three-celled sporophytic germling (right), each perched on an empty oogonium; (*g*) embryonic sporophyte with rhizoids.**

by the transition meristem. A few sporophytes are capable of almost indefinite growth.

Anatomy. Laminariales have a more complex anatomy than any other group of algae, although Fucales (rockweeds) nearly equal them in this regard. A mature stipe comprises three regions: a photosynthetic and meristematic thin outer layer (meristoderm); a parenchymatous cortex; and a filamentous medulla. The cortex is often pierced by mucilage ducts lined with small secretory cells. The medulla comprises both ordinary filaments and sieve tubes, the latter composed of superposed sieve elements with perforated end walls. Both types of filaments are interconnected by transverse filaments. Sieve tubes have been shown to translocate mannitol, a water-soluble sugar alcohol. Translocation is significant in large kelps such as *Macrocystis, Nereocystis,* and *Pelagophycus,* where photosynthate produced by blades floating on the surface is moved long distances to developing stipes and holdfasts at heavily shaded depths. The anatomy of the blade is similar to that of the stipe, except that the meristoderm is thinner and less active in cell division.

Life history and reproduction. The microscopic filamentous gametophytes, except in *Chorda tomentosa*, are dioecious. Female gametophytes (Fig. 1*f*) consist of relatively few stout cells and are sparsely branched. An egg, produced in an intercalary oogonium, is extruded through a rupture in the oogonial wall and remains attached to a mucilaginous cushion. The eggs of some species liberate a chemical that attracts sperm of the same species. Male gametophytes (Fig. 1*d*) have more cells, which are more slender, and are profusely branched. Antheridia, which are produced in clusters, each contain a biflagellate sperm (Fig. 1*e*). The zygote, attached to the empty oogonium, initiates a fairly regular sequence of cell divisions, resulting in an embryo that becomes anchored to the substrate by rhizoids (Fig. 1*g*). Eventually, the gametophyte is overgrown and the rhizoids are replaced by a permanent holdfast. The juvenile plant is clearly differentiated into holdfast, stipe, and blade when it is only 0.8–1.2 in.

(2–3 cm) high. Unilocular sporangia are produced in sori (Fig. 1*b*), accompanied by unicellular hairs (paraphyses), covering large portions of blades. The initial nucleus of a developing sporangium is diploid and undergoes meiosis followed by several mitotic divisions. Mature sporangia commonly contain 64 biflagellate zoospores (Fig. 1*c*), which germinate to produce gametophytes, half of them male and half female.

Classification. Four families of kelps are recognized. The sporophyte of *Chorda,* the only genus in the family Chordaceae, is unique among kelps in being whiplike and not differentiated into stipe and blade. Sporophytes of Laminariaceae display a simple plan of holdfast, stipe, and unbranched (but often laciniate) blade. Genera are distinguished by various patterns of ribs, folds, and perforations of the blade. *Laminaria,* which lacks such modifications, may be considered the basic genus of the family and perhaps of the entire order.

In sporophytes of Alariaceae, there is a differentiation between sterile and fertile blades, the latter (sporophylls) growing from the lateral margins of the blade, transition zone, or stipe. *Alaria* has a long sterile blade, which is ruffled and has a prominent midrib, and two clusters of short winglike sporophylls, one on either side of the transition zone. The long straplike stipes of *Egregia* (feather boa kelp) are fringed on either side with a mixture of sterile blades (some transformed into air bladders) and sporophylls.

In sporophytes of Lessoniaceae, the primary blade splits longitudinally, but unlike Laminariaceae, in which splitting affects only the blade, the splits extend into the transition zone. The meristem is thus dispersed, resulting in branching of the stipe, with each branch terminating in a blade. Sporophytes of some Lessoniaceae are essentially annual, with the thallus either torn from its substrate by winter storms or disintegrating in place; those of other Lessoniaceae are perennial and capable of nearly indefinite growth.

Nereocystis (bull kelp) is a spectacular plant of the Pacific coast of North America from Point Conception, California, northward. Its long, smooth, terete stipe expands into a large bulbous bladder just below the transition zone. Immediately above the bladder, the stipe branches dichotomously and produces numerous blades. The bladder, which buoys the blades at the surface, contains a mixture of gases including carbon monoxide. The sporophyte of the bull kelp appears in the spring and usually disappears in the winter, but it may persist without further growth into the second year. It reaches a length of at least 65 ft (20 m) within a few months, making it one of the world's fastest-growing plants. *Pelagophycus* (elk kelp), with antlerlike branches above the terminal bladder, is the counterpart of the bull kelp in southern California and northern Baja California.

The sporophyte of *Macrocystis* (giant kelp), which is perennial, is anchored by a massive cluster of haptera about 3.5 ft (1 m) broad. The numerous long branches may be regenerated from their base in response to abrasion, grazing, or harvesting. Each ultimate branch ends in a falcate blade, which splits progressively from base to apex to form a series of linear blades, each subtended by a bladder (**Fig. 2**). Sporophylls are distinct structures borne low on the major branches. *Macrocystis* equals or exceeds *Nereocystis* in rapidity of growth.

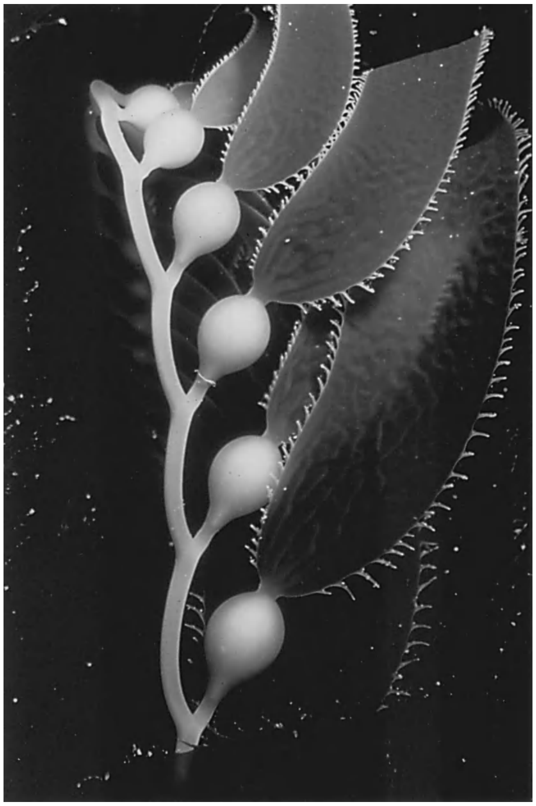


Fig. 2. ***Macrocystis*** **(giant kelp). Part of frond, showing series of blades, each subtended by a bladder. (*Underwater photo by Sylvia A. Earle*)**

Distribution. Kelps are largely confined to cold waters of both hemispheres, displaying a peak of diversity, abundance, and luxuriance all around the North Pacific. Only *Lessonia* and *Ecklonia* (both Lessoniaceae) are characteristic of the Southern Hemisphere, but *Macrocystis* is bipolar, occurring abundantly in subantarctic waters. Kelps are lacking in Antarctica, where their ecological niche is filled by members of the brown algal family Desmarestiaceae.

Some kelps are found in the intertidal zone, usually on rocks exposed to heavy waves, but most are subtidal. *Postelsia* (sea palm) shows amazing tenacity in facing the full force of the surf in the most highly exposed sites. *Macrocystis pyrifera* forms extensive subtidal stands regardless of exposure. The plants are usually anchored at depths of 20–65 ft (6–20 m), with their blades buoyed up toward or on the surface of the water. *Nereocystis* seeks slightly shallower water, often in protected coves.

Economic and ecological importance. Kelps were traditionally used by coastal farmers of the North

Atlantic as a wet manure, or were burned to an ash, which was then used as a fertilizer or in industry. The word kelp, in fact, was originally applied to the ash. During World War I, a potash industry flourished on the Pacific coast of the United States, using *Macrocystis* as the source. The principal use of kelps today is in the alginate industry, with *Macrocystis* the chief source in California, and *Laminaria* together with the rockweed *Ascophyllum* the chief source in the North Atlantic. Various kelps are used extensively for human food in Japan, including *Laminaria* (kombu) and *Undaria* (wakame). In China *Laminaria japonica* has been established as an important maricultural crop plant. Ecologically, such kelps as *Macrocystis* and *Nereocystis* are particularly important because they form protective canopies harboring special communities of smaller seaweeds, invertebrates, fishes, and the sea otter. *See* ALGAE; ALGINATE.

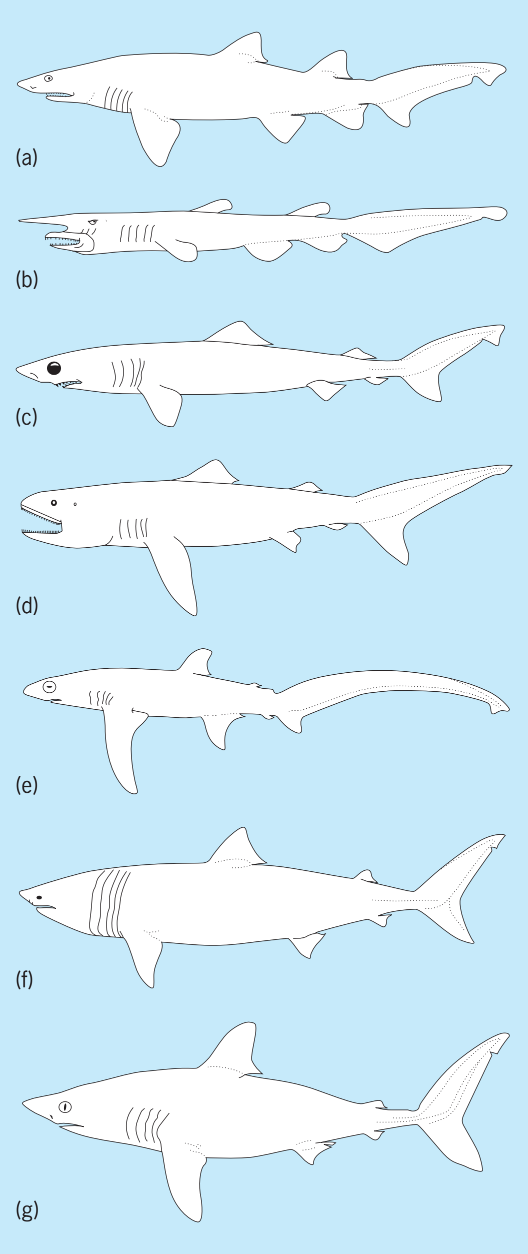
Paul C. Silva; Richard L. Moe

Bibliography. H. C. Bold and M. J. Wynne, *Introduction to the Algae: Structure and Reproduction*, 2d ed., 1997.

Lamniformes

An order of sharks distinguished from all other sharks by the following combination of characters: trunk cylindrical; head conical to cylindrical, not laterally expanded; two dorsal fins, lacking spines; five gill slits, all anterior to the pectoral fin base or the last two above the pectoral fin base; mouth large, its gape extending well beyond the eyes; spiracles usually present, small and well behind the eyes; eyes without a nictitating membrane; barbels absent; and ring-type intestinal valve. Lamniform sharks are ovoviviparous. Fertilized eggs are enclosed in egg cases in the oviducts (uteri), where they undergo embryonic development, ordinarily from the energy supplied by the yolk. Although the reproductive event may begin with each uterus containing about 20 fertilized eggs, normally only one or two in each uterus reach full term. The gross loss of potential progeny is explained by oophagy or uterine cannibalism, an uncommon phenomenon in which the largest (and presumably oldest) fetuses are free in the uterus to feed on the remaining eggs and developing embryos, thus cannibalizing their potential siblings. This reproductive strategy sacrifices numbers for a few large, strong pups that are like small adults, ready to fend for themselves. *See* ELASMOBRANCHII; SELACHII.

Lamniformes comprises only 15 extant species in seven families (see **illustration**), four of which are monotypic.



Representatives of seven Lamniformes families. (*a*) Odontaspidae. (*b*) Mitsukurinidae. (*c*) Pseudocarchariidae. (*d*) Megachasmidae. (*e*) Alopiidae. (*f*) Cetorhinidae. (*g*) Lamnidae. (*From J. S. Nelson, Fishes of the World, 4th ed., Wiley, New York, 2006*)

Odontaspidae (sand tiger sharks). Sand tiger sharks are distinguished by gill openings that are all in front of the pectoral fin; large teeth with slender cusps and lateral cusplets; a compressed caudal peduncle, without a lateral keel but with an upper precaudal pit; and an asymmetrical caudal fin with a relatively short ventral lobe. Females are the larger of the sexes, with a maximum total length of 370 cm (12 ft).

The family comprises two genera, *Carcharias* with one species and *Odontaspia* with two species. The snout is short and flattened in *Carcharias*, and long and conical in *Odontaspia*; dorsal, anal, and pelvic fins are about the same size in *Carcharias*, whereas the first dorsal is larger than the second, and the second is larger than the anal, in *Odontaspia*.

Odontaspids are known from tropical to cool-temperate zones of the Atlantic, Indian, and Pacific oceans, from continental and insular shores to the outer shelves and slopes to about 1600 m (5250 ft). They are voracious feeders on a wide variety of bony fishes, other sharks, rays, squids, and bottom crustaceans. Schools of sand tigers have been observed cooperatively surrounding and bunching schooling prey and then feeding on them. Of the three species, *Carcharias taurus* is by far the best known biologically. The gestation period is 8 to 9 months and the newly born are 95–105 cm (37–42 in.) in total length.

Relatively few attacks by these sharks on people have been recorded, but any large shark should be regarded as potentially dangerous and should not be harassed or provoked.

Mitsukurinidae (goblin sharks). This family consists of only one extant species, *Mitsukurina owstoni*, a bizarre shark that occurs in apparently disjunct populations in the eastern and western Atlantic (including the Gulf of Mexico), off South Africa, in the eastern Pacific (including southern California), and off southern Australia. It is easily recognized by an elongated and flattened snout; small eyes; soft flabby body and fins; very protrusible jaws; teeth with long slender cusps and no cusplets; dorsal fins about equal in size, and smaller than anal and pectoral fins; no precaudal pit; and caudal fin without a ventral lobe. Maximum length is 3.8 m (12.5 ft).

The goblin shark is an uncommon, poorly known, bottom-dwelling shark that inhabits outer continental shelves and beyond to at least 550 m (1800 ft), but rarely occurs near shore. Very little is known of its biology; however, its morphology suggests a relatively inactive, slow-swimming shark that probably uses its sensitive snout to detect prey and its highly protractile jaws to reach out and seize the prey, which may be squid, shrimp, crabs, or a variety of fishes.

Pseudocarchariidae (crocodile sharks). This is a monotypic family, represented by *Pseudocarcharias kamoharai*, which occurs in scattered localities, usually oceanic, in tropical and subtropical waters of the Northern and Southern hemispheres. It has a slender spindle-shaped body; long gill slits that reach dorsally past the axis of the body; highly protrusible jaws that are equipped with long slender teeth; angular dorsal fins; a prominent ventral caudal fin lobe; exceptionally large eyes; upper and lower precaudal pits; and a low lateral keel. This is the smallest lamniform shark, with a maximum adult length of 110 cm (43 in.); it is 41 cm (16 in.) at birth.

Megachasmidae (megamouth shark). This monotypic family, represented by *Megachasma pelagios*, was discovered in 1976 off Oahu, Hawaii. Since then, 35 more specimens (as of May 2006) have been captured in scattered places in warm temperate and tropical shallow coastal waters [depth: 5 m (16 ft)] to oceanic waters, where it inhabits the epipelagic zone (the illuminated surface zone) to depths of 166 m (545 ft). Megamouth is a very distinctive shark in having a large terminal mouth; large head; cylindrical and somewhat compressed trunk; snout that is exceptionally short but broadly rounded; moderately long gill slits, the last two over the pectoral fin base; upper and lower precaudal pits; small teeth in numerous rows; and gill rakers in the form of fingerlike dermal papillae. At a maximum length of 5.5 m (18 ft), megamouth is the smallest of the three giant filter-feeding sharks, the others being the basking shark [10 m (33 ft) or more] and the whale shark [12 m (39 ft) or more].

Alopiidae (thresher sharks). Thresher sharks are easily distinguished by the greatly elongate and curved upper caudal fin lobe, which is about half their total length; long narrow pectoral fins; relatively small gill slits, with the third to fifth over the origin of the pectoral fin; large eyes; a small mouth; precaudal pits; and a caudal peduncle without keels. Maximum length is 5.7 m (18.7 ft).

Thresher sharks occur practically circumglobally in temperate to tropical waters from coastal to oceanic regions. They feed on a variety of schooling fishes, such as mackerels and herrings, as well as squids, octopuses, and pelagic crustaceans. The long caudal fin is used to herd fishes and to slash through a school, killing or stunning prey to be consumed later. There is one genus with three species: *Alopias pelagicus* (pelagic thresher), 347 cm (11 ft) total adult length, 96 cm (38 in.) at birth; *A. superciliosus* (bigeye thresher), 488 cm (16 ft) total adult length, 105 cm (41 in.) at birth; and *A. vulpinus* (thresher shark), 760 cm (25 ft) total adult length, 150 cm (59 in.) at birth.

Cetorhinidae (basking sharks). Cetorhinidae is another monotypic family, which is represented by the basking shark (*Cetorhinus maximus*). It occurs on continental and insular shelves, offshore and often close to land, just free of the surf zone, in boreal (northern) to warm temperate waters of the Atlantic and Pacific oceans. It is easily distinguished from other sharks by the following combination of characters: exceptionally large gill openings practically encircling the head and all in front of the pectoral fin; hairlike gill rakers of modified dermal denticles on the gill arches; very small and hooklike teeth; nearly lunate caudal fin, with a well-developed lower lobe; precaudal pits; and caudal peduncle with strong lateral keels.

The basking shark is a filter feeder, but its method differs from the other two huge filter-feeding sharks (megamouth and whale shark), as it passively strains plankton while swimming with mouth agape and external gill slits open, as opposed to pumping or gulping water into the pharynx and force-expelling it across the gill rakers. The basking shark is the world's second largest living fish, reportedly reaching 15 m (49 ft) in total length, but specimens exceeding 10 m (33 ft) are exceptional. Size at birth is unknown; however, the smallest known free-living basking shark measured 165 cm (65 in.).

Lamnidae (mackerel sharks). Mackerel sharks occur virtually globally from nearshore and the continental shelf to epipelagic zones of cold-temperate to tropical waters. They are distinguished from other sharks by the following combination of characters: fusiform body moderately slender to robust; snout moderately long, pointed, and conical; large gill openings, extending above the axis of the body (except in *Lamna*) and all in front of the pectoral fin base; no gill rakers on the gill araches; first dorsal fin large, high, erect, and angular or somewhat rounded; second dorsal and anal fins very small; caudal fin lunate or crescent-shaped and caudal peduncle with distinct keels. These are fast-swimming predators with a high level of activity made possible by partial warm-bloodedness (sustained activity is by warm musculature). The family comprises three genera and five living species.

Carcharodon carcharias (white shark or great white shark) inhabits coastal and offshore continental and insular waters and occasionally shallow bays. It is one of the most widely ranging sharks, although uncommon in the tropical waters. Certainly, it is the most notorious and most feared shark. Its large size, large teeth, powerful jaws, and powerful and efficient scombrid-like mode of locomotion that allows it to sustain long cruises make this shark a formidable predator. It feeds on a wide variety of bony fishes, as well as on sharks, rays, seals, dolphins and porpoises, sea birds, squid, octopi, crabs, and carrion. Although its role as a man-eater has been exaggerated, the white shark is still considered very dangerous.

The white shark probably does not exceed 7.5 m (25 ft) in total length. Captured individuals are commonly between 1.4 and 6 m (4.5 and 20 ft). Even at 6 m, it is the world's largest predator. The famous late Pliocene fossil, *Carcharodon megalodon*, known commonly as megatooth shark or megalodon, which reached lengths up to 20 m (66 ft), is a member of this family.

Isurus oxyrinchus (shortfin mako) is a coastal and oceanic inhabitant of the temperate and tropic seas of the world. It is probably the fastest of all sharks, capable of extreme bursts of speed when hooked or in pursuit of prey, and can leap out of the water to a height several times its length. Its maximum total length is about 400 cm (13 ft).

Isurus paucus (longfin mako) is a little-known circumglobal epipelagic shark in the warm-temperate and tropical waters. It is not uncommon in the western North Atlantic and places in the mid-Pacific, but it is rather rare elsewhere. The extremely long pectoral fins suggest a less active shark than its congener, the shortfin mako. Its maximum total length is 417 cm (14 ft).

Lamna ditropis (salmon shark) is the most robust of the lamnid sharks. It is a coastal and oceanic inhabitant of the North Pacific from Japan and Korea to the Bering Sea, and southward to southern California. It is a swift swimmer and voracious predator, and maintains a body temperature several degrees above ambient water temperature. It preys heavily on salmon and attains a maximum total length of about 305 cm (10 ft).

Lamna nasus (porbeagle) is a epipelagic shark that occurs in the western North Atlantic from New Jersey to the Gulf of St. Lawrence, across the Atlantic to the Mediterranean Sea and northward to the western Barents Sea. It also occurs in the cold temperate waters of the Southern Hemisphere, primarily in the southern Indian Ocean, off southern Australia and New Zealand, and off southern South America, tending to avoid water temperatures greater then 18°C (65°F). The porbeagle is most common on offshore fishing banks, where it feeds on a variety of bony fishes, as well as on other sharks and squids. Maximum total length is no less than 300 cm (9.8 ft), possibly 370 cm (12 ft), and it is 60–75 cm (24–30 in.) at birth.

Herbert Boschung

Bibliography. H. B. Bigelow and W. C. Schroeder, Sharks, in Fishes of the Western North Atlantic, *Mem. Sears Found. Mar. Res. Mem.*, 1(pt. 1):29–556, 1948; L. J. V. Compagno, Checklist of Chondrichthyes, pp. 503–547, in W. C. Hamlett (ed.), *Reproductive Biology and Phylogeny of Chondrichthyes: Sharks, Batoids and Chimaeras*, Science Publishers, Enfield, NH, 2005; L. J. V. Compagno, Checklist of living elasmobranches, pp. 471–498, in W. C. Hamlett (ed.), *Sharks, Skates, and Rays: The Biology of Elasmobranch Fishes*, Johns Hopkins University Press, Baltimore, 1999; L. J. V. Compagno, *FAO Species Catalogue*, vol. 4: *Sharks of the World: An Annotated and Illustrated Catalogue of Shark Species Known to Date, Part 1: Hexanchiformes to Lamniformes*, FAO Fish. Synop. 125, vol. 4, pt. 1:212–249, 1984; L. J. V. Compagno, *Sharks of the Order Carcharhiniformes*, Princeton University Press, 1988; L. J. V. Compagno, *Sharks of the World: An Annotated and Illustrated Catalogue of Shark Species Known to Date, FAO Species Catalogue for Fishery Purposes* no. 1, vol. 2: *Bullhead, Mackerel and Carpet Sharks (Heterodontiformes, Lamniformes and Orectolobiformes*), FAO, Rome, 2001; J. S. Nelson, *Fishes of the World*, 4th ed., Wiley, New York, 2006.

Lamp

A generic term for a device designed to produce electromagnetic radiation in the form of light, heat, ultraviolet energy, or a combination of the three. The term lamp is applied to the entire range of sources, including flame sources (such as kerosine lamps or gas lamps with Welsbach mantles), incandescent sources, and electric arc discharge sources. Used with a modifier, such as ultraviolet, infrared, or sun, the term lamp is used to indicate sources that radiate energy in the ultraviolet or infrared portions of the electromagnetic spectrum (plus some radiation in the visible part of the spectrum).

Electric lamps are the most common and useful types of light-producing sources. These devices convert electrical energy into light (radiant energy in the visible portion of the spectrum). The most widely used lamps for lighting buildings and areas, such as parking lots, stadia, and streets, are incandescent lamps, fluorescent lamps, and high-intensity arc discharge lamps (HID lamps). *See* ARC LAMP; FLUORESCENT LAMP; INCANDESCENT LAMP; INFRARED LAMP; SUN; ULTRAVIOLET LAMP; VAPOR LAMP.

G. R. Peirce; Robert Leroy Smith

Bibliography. Illuminating Engineering Society, *IES Lighting Handbook*, 8th ed., 1993.

Lampriformes

An order of fishes also known as Lampridiformes and Allotriognathi. Although the families of Lampriformes appear very dissimilar, they share the following combination of characters that distinguish them from other orders of actinopterygian (ray-finned) fishes: notably compressed body, often ribbonlike; no true spines in fins, rays only; thoracic pelvic fins, if present; swim bladder, if present, physoclistous; and small cycloid scales, commonly absent. In addition, lampriforms have a unique type of protrusible upper jaw, by which the premaxilla excludes the maxilla from the gape, and the maxilla, instead of having a ligamentous attachment to the ethmoid and palatine bones, slides in and out with the highly protractile premaxilla. Most species are large and colorful; some have inspired sea-serpent stories. *See* ACTINOPTERYGII; SWIM BLADDER.

Strong evidence has been presented by several authors supporting the notion that lampriforms are not percomorphs but prepercomorphs and are the primitive sister group to the remaining Acanthomorpha (spiny-rayed fishes). Dating from the Cretaceous Period (135 million years before present), lampriform fishes are presently represented by 21 species in 12 genera and seven families.

Two families (Veliferidae and Lampridae), collectively known as bathysomes, have deep compressed bodies with a symmetrical caudal fin and a well-developed skeleton; the remaining five families (Stylephoridae, Lophotidae, Radiicephalidae, Trachipteridae, and Regalecidae), called taeniosomes, have long ribbonlike bodies with an asymmetrical caudal fin and a weak skeleton.

Veliferidae (velifers). These fish are found in the Indian and western mid-Pacific oceans, usually near the shore. Dorsal and anal fins long and high, with 32–44 and 25–35 rays, respectively; lateral line follows dorsal contour of the body; teeth absent; swim bladder is bifurcate and extends posteriorly far beyond anus; and six branchiostegal rays. Maximum length about 40 cm (16 in.). Two monotypic genera, *Velifer* and *Metavelifer* (**Fig. 1**).

Fig. 1. Spinyfin velifer (*Metavelifer multiradiatus*). (*Photo by Jack E. Randall*)

Lampridae (opahs). The opahs are pelagic, living in the open Atlantic, Indian, and Pacific oceans. The dorsal fin is falcate (sickle-shaped) anteriorly; otherwise, it and the anal fins are long and low, with 48–56 and 33–42 rays, respectively; the lateral line is arched high in front; the pectoral fins are falcate; and minute cycloid scales are present. Maximum length is up to 180 cm (71 in.). Two species, *Lampris guttatus* (worldwide) and *L. immaculata* (cold and temperate waters of the Southern Hemisphere), are known.

Stylephoridae (tube-eyes or thread-tails). These are oceanic fishes, occurring at depths of 300–800 m (980–2620 ft). The body is slender and ribbonlike with up to 200 vertebrae; the dorsal fin extends from the nape to the tail, with 115–124 rays; the anal fin is short, with 15–17 rays; the pelvic fin has only one ray; the caudal fin has seven rays, with the lowermost two being extremely elongate; eyes are large, telescopic, and directed forward or upward; the mouth is small; the snout is tubular; teeth are small; and there is no swim bladder. The only described species, *Stylephorus chordatus*, swims in a vertical position, head uppermost, and feeds by rapidly expanding its buccal cavity, effectively sucking in plankton. Maximum length is 31 cm (12 in.).

Lophotidae (crestfishes). These are oceanic fishes. The dorsal fin originates above or before the tip of the snout and continues to the caudal fin, with 220 to 392 rays; the anal fin is short and posteriorly placed; the body has small deciduous cycloid scales and may appear naked; the caudal fin is symmetrical or nearly so; the pelvic fins are present or absent; the swim bladder is present; and the ink sac opens into the cloaca, with the ink possibly serving to blind predators. Two genera and three species have been identified. Maximum length is about 200 cm (80 in.).

Radiicephalidae (tapertails). These fish are found in the central and eastern Atlantic, off New Guinea, and in the California Current region. The body tapers to a thin caudal filament; the caudal fin has a small upper lobe of four rays and a long slender lower lobe of seven rays; the dorsal fin possesses 152–159 rays; the anal fin is vestigial; scales are found along the lateral line, with the remainder of the body naked; the swim bladder is well developed; the brown ink sac discharges into the cloaca; and there are 114–121 vertebrae. One species, *Radiicephalus elongates*,

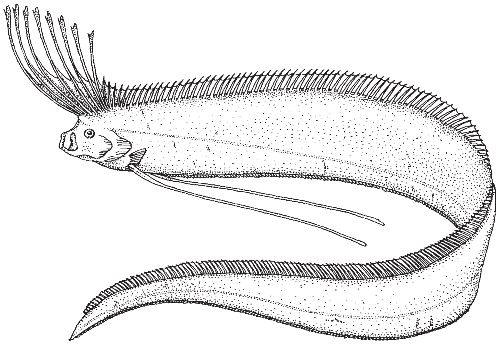


Fig. 2. Oarfish (*Regalecus glesne*). (*After D. S. Jordan and B. W. Evermann, The Fishes of North and Middle America, U.S. Nat. Mus. Bull. 47, 1900*)

is known, although it is rarely captured. Maximum length is about 70 cm (28 in.).

Trachipteridae (ribbonfishes). The ribbonfishes are bathypelagic [living at ocean depths between 600 and 3600 m (2000 and 12,000 ft)] in the Arctic, Atlantic, Indian, and Pacific oceans, as well as the Mediterranean Sea. The body is naked or covered with deciduous cycloid scales or with deciduous modified ctenoid scales; there is no anal fin; the caudal fin is long and at a right angle to the body, usually consisting of an upper lobe only; the dorsal fin is very long, originating distinctly behind the tip of the snout; the eyes are large; teeth are present; the swim bladder is rudimentary or absent; and there are 62–111 vertebrae. Three genera (*Desmodema*, *Trachipterus*, and *Zu*) and 10 species are known. Maximum length is about 170 cm (67 in.).

Regalecidae (oarfishes). Oarfishes are found worldwide in warm-temperate seas. The body is very elongate; scales are absent; the skin is covered with tubercles; the dorsal fin is long (260–412 rays), beginning above the eye and continuing to nearly the end of the tail; the front rays are very long; the pelvic fins are long and slender, with one ray, having a paddlelike expansion near the tip; the anal fin, teeth, and swim bladder are absent; the eyes are small; and there are about 143–170 vertebrae. Two species, *Regalecus glesne* and *Agrostichthys parkeri*, have been identified.

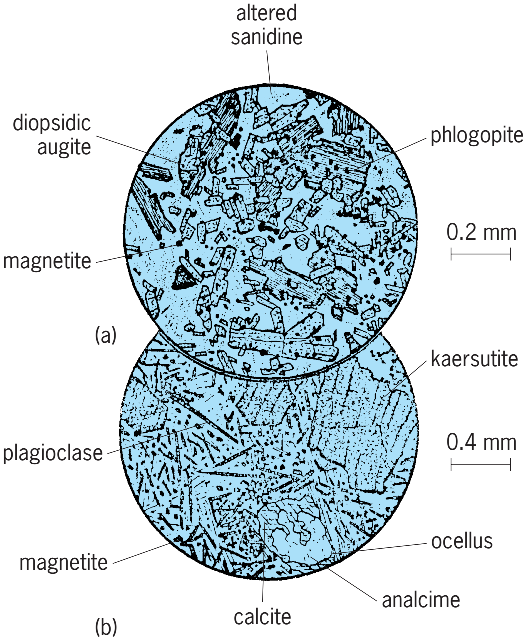
Regalecus glesne (**Fig. 2**) is called oarfish locally, but king-of-the-herring in Europe, from an old tale of this fish announcing the arrival of herrings. The oarfish is a spectacular creature, having a huge ribbonlike body, bright red dorsal fin (the front part crowning the head with long plumelike rays), and long bright-red pelvic fins. Attaining a maximum length of 800 cm (26 ft), it is the longest bony fish and undoubtedly the source of many sea-serpent stories.

Herbert Boschung

Bibliography. S. R. Charter and H. G. Moser, Radiicephalidae: Tapertails, pp. 665–667 in H. G. Moser (ed.), The early stages of fishes in the California Current region, *California Cooperative Oceanic Fisheries Investigations (CalCOFI) Atlas*, no. 33, 1996; G. D. Johnson and C. Patterson, Percomorph phylogeny: A survey of acanthomorphs and a new proposal, *Bull. Mar. Sci.*, 52(1):554–626, 1993; J. S. Nelson, *Fishes of the World*, 4th ed., Wiley, New York, 2006; J. E. Olney, Lampriformes: Development and relationships, in H. G. Moser et al. (eds.), *Ontogeny and Systematics of Fishes*, Spec. Publ. no. 1, American Society of Ichthyology and Herpetology, 368–379, 1984; J. E. Olney, G. D. Johnson, and C. C. Baldwin, Phylogeny of lampridiform fishes, *Bull. Mar. Sci.*, 52(1):137–169, 1993; C. R. Robins and G. C. Ray, *A Field Guide to Atlantic Coast Fishes of North America*, Haughton Mifflin, Boston, 1986; M. M. Smith and P. C. Heemstra (eds.), *Smith's Seafishes*, Macmillan South Africa, Johannesburg, 1986; V. Walters, Synopsis of the lampridiform suborder Veliferoidei, *Copeia*, 1960(3):245–247, 1960; E. O. Wiley, G. D. Johnson, and W. W. Dimmick, The phylogenetic relationships of lampridiform fishes (Teleostei: Acanthomorpha), based on a total evidence analysis of morphological and molecular data, *Mol. Phylogenet. Evol.*, 10(3):417–425, 1998.

Lamprophyre

Any of a heterogeneous group of gray to black, mafic igneous rocks characterized by a distinctive panidiomorphic and porphyritic texture in which abundant euhedral, dark-colored ferromagnesian (femic) minerals (dark mica, amphibole, pyroxene, olivine) occur in two generations—both early as phenocrysts and later in the matrix or groundmass—while felsic minerals (potassium feldspar, plagioclase, analcime, melilite) are restricted to the groundmass (see **illus.**). Compared with the common igneous rocks,



Specimens of lamprophyre. (*a*) Minette, found near Walsenburg, Colorado. The altered sanidine matrix also contains other minor minerals. (*b*) Camptonite, found in Campton Falls, New Hampshire. (*After H. Williams and C. M. Gilbert, Petrography: An Introduction to the Study of Rocks in Thin Sections, W. H. Freeman, 1954*)

Common lamprophyres

Dominant or diagnostic femic mineral	Dominant or diagnostic felsic mineral			
	K-feldspar	Plagioclase	Analcime or glass	Melilite
Phlogopite	Minette	Kersantite	—	Alnoite
Hornblende	Vogesite	Spessartite	—	—
Kaersutite	—	Camptonite	Monchiquite	—

lamprophyres are chemically peculiar. They have low silicon, moderate aluminum, and variable calcium contents, but are rich in alkalies (potassium or sodium or both), magnesium, iron, and volatile constituents (H_2O and CO_2), and contain a wide variety of such minor and trace elements as titanium, phosphorus, barium, strontium, rubidium, zirconium, lanthanum, uranium, thorium, chromium, nickel, and cobalt. Many lamprophyres contain ocelli, small spheroidal bodies consisting of alkali-rich feldspar, analcime, or calcite with minor femic minerals.

Mineralogy and terminology. Many varieties of lamprophyre are known; only the more abundant ones are listed in the **table**. Minettes, kersantites, vogesites, and spessartites are the most common and are sometimes collectively called calcalkaline lamprophyres. Camptonites and monchiquites are less common; alnoites are rare. These varieties, along with some others, are referred to as alkaline lamprophyres. Although lamprophyres are named on the basis of their dominant or diagnostic femic and felsic minerals, their terminology is not straightforward because there is much gradation in mineral assemblages, especially in number and abundance of femic minerals, and felsic minerals are often too altered to permit identification.

All lamprophyres contain essential hydrous minerals, such as dark mica and amphibole, frequently accompanied by significant pyroxene and less olivine. The mica is a magnesium-rich titanium-bearing biotite or phlogopite. Under the microscope the biotite in a thin section of a lamprophyre can be seen as pseudo hexagonal plates that are chemically zoned. The pale magnesium-rich centers of the crystals are surrounded by reddish-brown iron-rich rims; groundmass biotites are chemically similar to phenocryst rims. Amphibole is less common than biotite; it forms prisms or needles of a green or brown hornblende in calcalkaline lamprophyres, and a deep brown kaersutite in alkaline lamprophyres. Pyroxene occurs as stubby monoclinic crystals of diopsidic augite or titanaugite, frequently zoned. Orthorhombic pyroxene is absent. Magnesium-rich olivine is usually the least abundant femic mineral; unlike the biotite, amphibole, and pyroxene, it generally occurs only as phenocrysts and frequently is altered and pseudomorphed by calcite and other secondary minerals such as chlorite and members of the montmorillonite group.

Although in hand specimen the groundmass appears dark, a large proportion of it is composed of colorless plates or laths of potassium feldspar (sanidine or orthoclase) or plagioclase feldspar (albite-andesine in calcalkaline lamprophyres, andesine-labradorite in alkaline). In the rare feldspar-free varieties, monchiquite and alnoite, the groundmass consists predominantly of an alkali-rich commonly altered glass ± analcime, or melilite and a calcium carbonate. Apatite, a magnetitelike spinel, and calcite (primary or secondary or both) are ubiquitous accessory minerals in all lamprophyres; other minerals that may occur include sphene, quartz, analcime, zeolites, nepheline, haüyne, monticellite, pyrite, perovskite, and ilmenite (many of these may be secondary).

Occurrence. Lamprophyres are widespread but volumetrically minor rocks that apparently are restricted to the continents and are the last manifestation of igneous activity in a given area. They usually occur as subparallel or radial swarms of thin (~1.6–160 ft or 0.5–50 m) dikes or, less commonly, sills, volcanic neck fillings, or diatremes, or, rarely, lava flows (lamprophyric lavas are sometimes called lamproites). A swarm may consist of one or more varieties of lamprophyre, but calcalkaline and alkaline lamprophyres usually do not occur together. Calcalkaline varieties frequently are found in the region of large granodiorite or granite bodies, while alkaline lamprophyres tend to be associated with alkaline basalts, nepheline syenites, carbonatites, or kimberlites; many exceptions to this generalization are known.

Alkaline lamprophyres and the rare minette may bear xenocrysts and xenoliths, foreign crystals and rock fragments representing the material that composes the Earth's mantle under the continents at depths greater than approximately 47 mi (75 km). Xenoliths of the continental crust can be found in all lamprophyres. *See* PHENOCRYST; XENOLITH.

Origin. Like all igneous rocks, a lamprophyre is the product of the crystallization of magma (molten rock). The ultimate origin of lamprophyre magmas is the upper portion of the Earth's subcontinental mantle where localized heating causes partial melting (anatexis) of mantle rock. The exact composition of the liquid or magma thus formed, which differs from that of the original unmelted mantle rock (source rock), depends on several factors, including the temperature and pressure at which anatexis occurs, the composition of the source rocks, and the degree of partial melting. Since magma is less dense than solid rock, it rises upward, invades the overlying continental crust, and eventually solidifies. During its ascent from the mantle to its site of final emplacement, a magma's composition may be changed by several processes, two of the most important being assimilation of foreign material, especially portions of the rocks making up the continental crust; and partial (fractional) crystallization of the magma, resulting in removal of early-formed crystals of femic minerals because of density differences between these crystals and the magma from which they were formed.

The controversy surrounding the origin of lamprophyres centers on the chemical composition of the mantle-derived magma and the extent to which it has been modified prior to its final solidification. Some geologists maintain that a normal basaltic magma (the most common product of anatexis of the mantle) is made lamprophyric by a combination of the magma composition-changing processes. But continuing studies of the major- and trace-element chemistry of lamprophyres is indicating increasingly that many minettes and alkaline lamprophyres result from the crystallization of nearly unmodified mantle-derived magmas. This implies that magmas of lamprophyric composition can be generated within the upper mantle. Although lamprophyric magmas are believed to form by small degrees of partial melting of mantle rock of unusual chemical and mineralogical composition, details of the origin of these magmas are unclear. By the operation of processes such as assimilation of foreign material and partial magma crystallization, lamprophyric magmas may themselves change or evolve during their ascent, and this probably accounts for the gradations between the varieties. The origin of kersantites, vogesites, and spessartites is still a matter of debate; they may represent modified minettelike or other lamprophyric magma. *See* IGNEOUS ROCKS; MAGMA. Sharon W. Bachinski

Land drainage (agriculture)

The removal of water from the surface of the land and the control of the shallow ground water table improves the soil as a medium for plant growth. The sources of excess water may be precipitation, snowmelt, irrigation water, overland flow or underground seepage from adjacent areas, artesian flow from deep aquifers, floodwater from channels, or water applied for such special purposes as leaching salts from the soil or for achieving temperature control.

The purpose of agricultural drainage can be summed up as the improvement of soil water conditions to enhance agricultural use of the land. Such enhancement may come about by direct effects on crop growth, by improving the efficiency of farming operations or, under irrigated conditions, by maintaining or establishing a favorable salt regime. Drainage systems are engineering structures that remove water according to the principles of soil physics and hydraulics. The consequences of drainage, however, may also include a change in the quality of the drainage water.

Agricultural drainage is divided into two broad classes: surface and subsurface. Some installations serve both purposes.

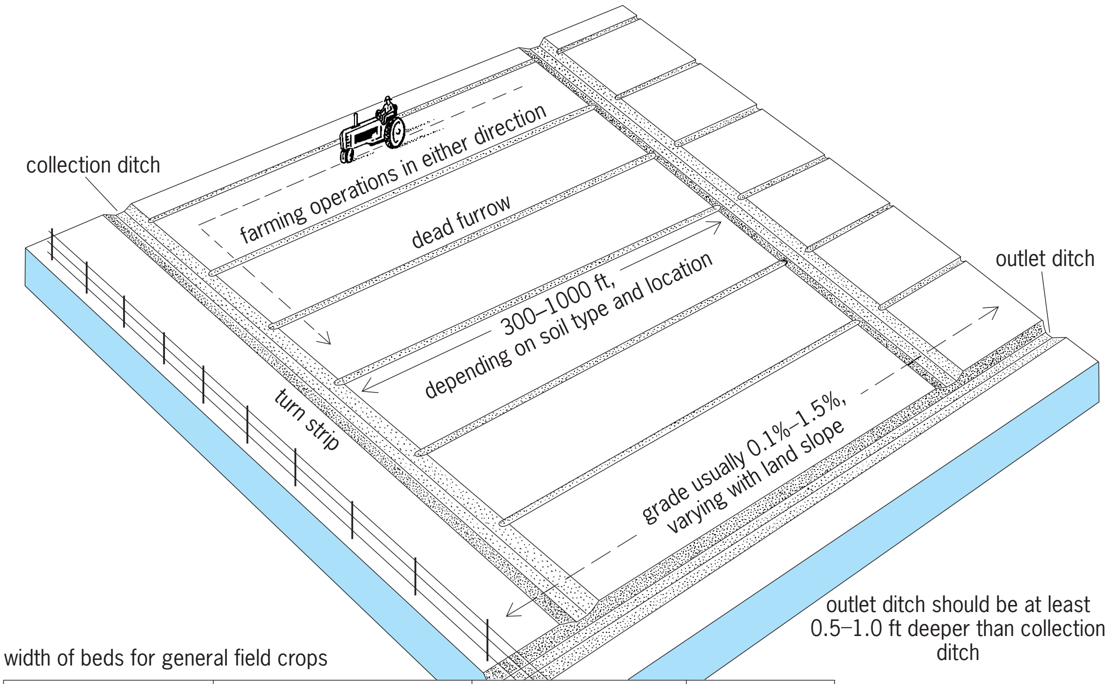
Surface drainage. Poor surface drainage conditions exist over large areas of land in the eastern half of the United States and Canada. The condition is caused by the inability of excessive rainfall to move over the ground surface to an outlet or through the soil to a subsurface drainage system. These poor surface drainage conditions are usually associated with soils that have low hydraulic conductivity. Often the soils are very shallow over a barrier such as rock or a very dense clay pan. The impermeable subsoil prevents the water from moving downward and prevents the proper functioning of a subsurface drainage system. Often the land slope is not sufficient to permit the water to flow across the ground surface. In other cases, the areas lack adequate drainage outlets. In order to correct this problem, something must be done to eliminate the depressions and to provide sufficient slope for overland flow. In addition, it is necessary to provide channels to convey the water from the affected area.

The practice of surface drainage may be defined as the diversion or orderly removal of excess water from the land surface by means of improved natural or constructed channels. The channels may have to be supplemented by shaping and grading the land surface so that the water may flow freely into the channel.

In some instances, a subsurface drainage system of pipes is needed in conjunction with surface drains. The effectiveness of the subsurface drains is increased by the removal of the water from the soil surface. As soon as the surface water is removed, the drain pipes can act to lower the water table and to provide a satisfactory environment for the growth of plants. There are essentially five types of surface field drainage systems in common use today: bedding system (**Fig. 1**); random ditch system; interception system; diversion ditch system; and field ditch system. Combinations of two or more systems may be required by circumstances encountered in the field. The choice of a particular system used for surface drainage depends upon the soil type, topography, crops to be grown, and farmer preference.

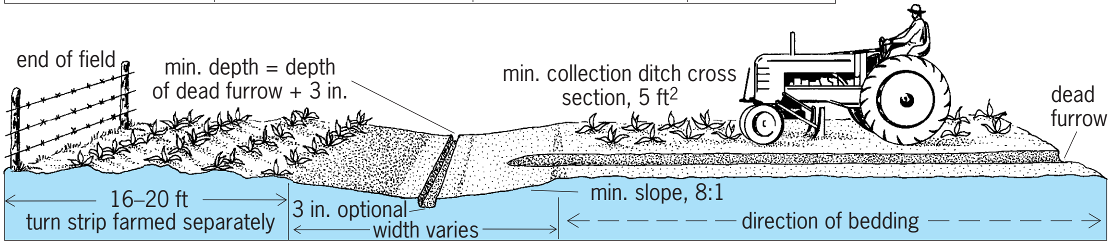
Subsurface drainage in humid regions. Subsurface drainage is required where a high water table is present. The main purpose of the drainage is to provide a root environment that is suitable for the maximum growth of plants and to sustain yields over long periods of time. One of the main reasons that poor drainage causes a decrease in crop production is the fact that the plant roots have only a limited amount of soil in which to grow. Not only do the plants lack food, but the plant roots suffer from a deficiency of oxygen which is needed for the respiratory processes, as the water that fills the soil pores displaces the air in the soil. Moreover, the water obstructs the gases which are given off by the roots, and some of these gases inhibit plant growth.

The critical need for drainage occurs in the early spring months when the plants are germinating. Lack of drainage retards the normal rise in the soil temperature, decreases plant resistance to disease, and inhibits root development. Poor drainage discourages the growth of aerobic bacteria which are needed to

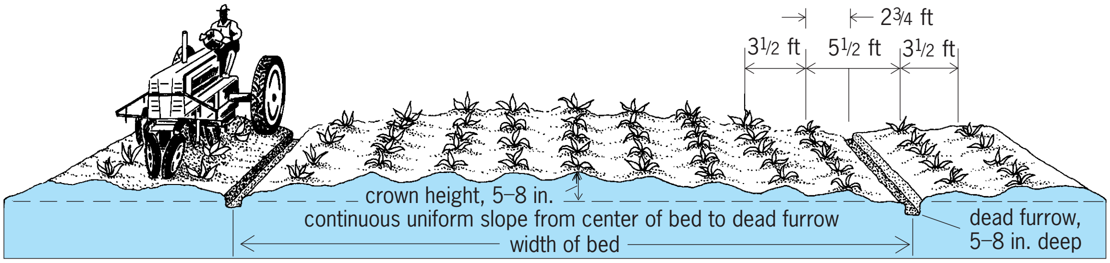


width of beds for general field crops

degree of internal drainage of the soil	width of bed in feet, center to center of dead furrows	number of 3½-ft corn rows with 2 ft allowed per dead furrow	number of rounds using 2–14 in. plows
very slow	23	6	5
	30	8	6½
	37	10	8
slow	44	12	9½
	51	14	11
fair	58	16	12½
	65	18	14
	72	20	15½
	79	22	17
	86	24	18½
	93	26	20



cross section at end of field showing collection ditch and turn strip



cross section of bed showing crown effect and proper pacing of corn rows

Fig. 1. Surface drainage bedding system. 1 in. = 25 mm; 1 ft = 0.3 m; 1 ft^2 = 0.09 m^2. (*USDA*)

supply nitrogen for crops. Toxic organic and inorganic compounds develop in saturated soils.

The depth of drains in humid regions is largely determined by soil conditions. Drain depths from 24 to 48 in. (60 to 120 cm) are commonly used to control a shallow ground water table. The spacing of the drains depends upon the soil hydraulic conductivity and the amount of the rainfall that must be removed.

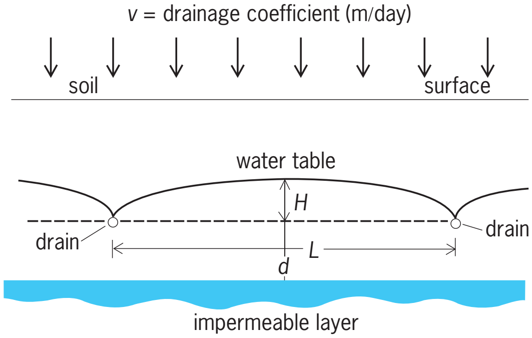


Fig. 2. Hooghoudt's drain spacing formula. Symbols are explained in the text.

Spacings vary between 33 and 100 ft (10 and 30 m).

A number of drain spacing formulas have been proposed and one of the most successful is S. B. Hooghoudt's formula, Eq. (1), where L = drain

$$L^2 = \frac{4K}{\nu}(H^2 + 2dH) \qquad (1)$$

spacing (meters), K = hydraulic conductivity (meters/day), ν = drainage coefficient, which is the rate of removal of water from the soil that is necessary to protect the crop from damage (meters/day), H = height of water table above the plant through drains (meters), d = distance from the plane through drains to the impermeable layer. Developed for the drainage of land in the Netherlands, the formula is based upon steady-state replenishment of the ground water. The height of the water table halfway between the drains is calculated as a function of the soil hydraulic conductivity and the rate at which water replenished the ground water table. The significance of the various parameters involved in the formula are presented in **Fig. 2**.

Drainage coefficients are of the order of 5–10 mm/day in Europe. Somewhat higher values (1–4 cm/day) are used in the United States.

Subsurface drainage in arid regions. Irrigation waters contain substantial quantities of salt, from 6 to 240 lb per 1000 ft^3 (0.1 to 4 metric tons per 1000 m^3). Irrigation water is applied at rates of 140,000 to 210,000 ft^3 per acre per year (10,000 to 15,000 m^3 per hectare per year) and hence between 0.05 to 26 tons is added to each acre annually (0.1 to 60 metric tons per hectare). Some of the applied salt precipitates in the soil; a small proportion is used by the plants, and the remainder must be removed from the soil by adding an amount of irrigation water in excess of the crop needs.

The output of salt in the drainage water must equal the input of salt in the irrigation water. If precipitation of salts in the soil and plant uptake of salts are ignored, it can be stated that salt input equals salt output.

The leaching requirement (LR) is defined as the fraction of irrigation water that must be drained in order to maintain the salt balance, as in Eq. (2),

$$\mathrm{LR} = \frac{\nu}{I} \qquad (2)$$

where ν represents the drainage coefficient and I the amount of applied irrigation water. The computation of ν depends on the salt tolerance of the crop to be grown.

In arid regions, it is necessary to control the ground water table well below the plant root zone. The plants will extract water from the soil, leaving salt behind. This results in a concentration of salts in the plant root zone; therefore, the water table must be maintained well below the plant root zone so that capillary rise of this salty ground water into the plant root zone is reduced. The depth of the drains then is determined by capillary rise into the plant root zone. Normally drains in irrigated areas are placed about 6 ft (1.8 m) deep. The spacing between the drains may be determined by Hooghoudt's formula.

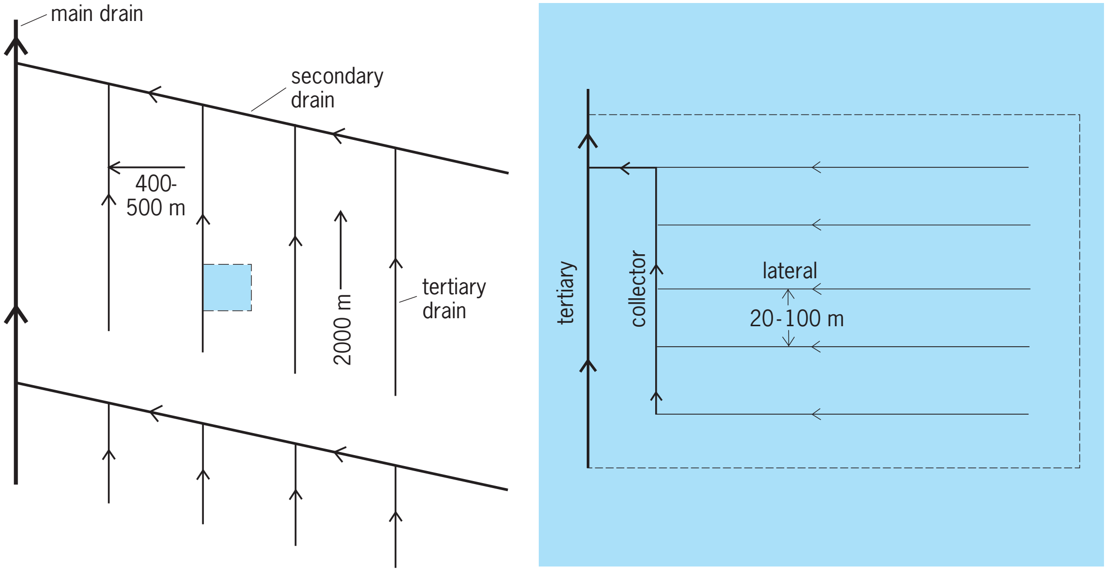


Fig. 3. Typical drainage system for an irrigation project. Shaded box on left is shown in detail on right. 1 m = 3.3 ft.

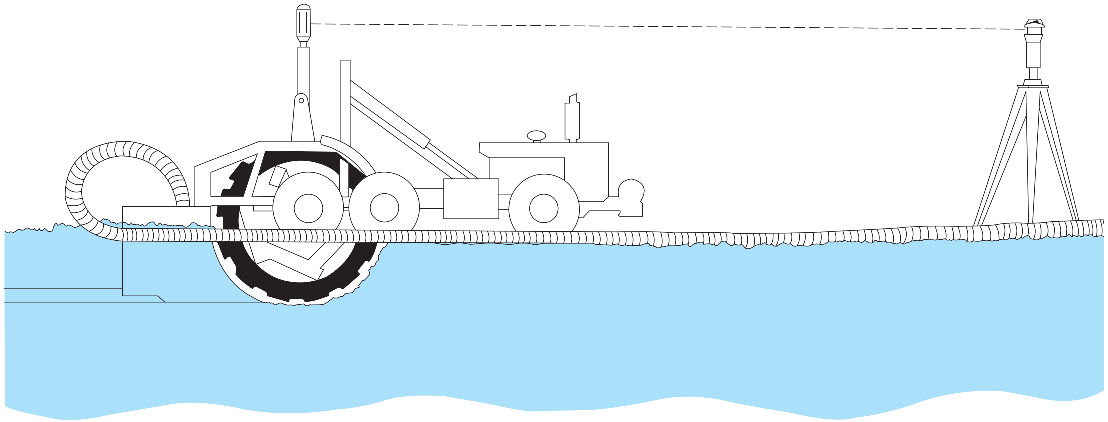


Fig. 4. Grade control by laser beam.

However, the drainage coefficient is determined by a consideration of the leaching requirement, precipitation of salts in soil from irrigation water, amount of leaching due to winter rainfall, and salt tolerance of crops. Typical drain spacings range from 66 to 330 ft or 20 to 100 m (**Fig. 3**).

In addition to the steady-state formula of Hooghoudt, analyses have been made of the transient water table situation. The U.S. Bureau of Reclamation utilizes this transient water table formula for determining the average depth of water table during the growing season. If according to the bureau's method of analysis the water table is rising during the growing season, and continues to rise, the condition indicates a serious drainage problem will occur. The bureau's drainage criteria then are based upon controlling the water table so that it either remains constant or declines over a period. The procedure is called the dynamic equilibrium method.

Construction of drains. The construction of subsurface drains was revolutionized in the years 1976–1979. In the past, a variety of materials such as rocks, clay pipe, concrete pipe, and other materials were used for subsurface drains. With the invention of machines that make perforated corrugated plastic pipe, this pipe largely supplanted the other materials. Corrugated plastic pipe has the advantage of light weight, and it comes in long lengths so the cost of handling is reduced.

Drain pipes either are laid in the bottom of a trench or are pulled into place by a drain plow. The development of laser beams for grade control permits the use of drain plows (called the trenchless method) for installing corrugated plastic tubing (**Fig. 4**).

In unstable soils usually found in arid regions, it is necessary to surround the drain pipe with gravel, sand, synthetic fibers, or organic material in order to prevent fine sands and silts from entering the pipe with the drainage water. *See* AGRICULTURAL SOIL AND CROP PRACTICES; IRRIGATION (AGRICULTURE); PLANT-WATER RELATIONS.

J. N. Luthin

Bibliography. G. R. Bumli (ed.), *Principles of Project Formulation for Irrigation and Drainage Projects*, 1982; A. Dinar, *The Economics and Management of Water and Drainage in Agriculture*, 1991; S. C. Harris (ed.), *Irrigation and Drainage*, 1990; E. G. Kruse and C. R. Burdick (eds.), *Environmentally Sound Water and Soil Management*, 1982; J. N. Luthin (ed.), *Drainage Engineering*, rev. ed., 1978.

Land reclamation

The process by which seriously disturbed land surfaces are stabilized against the hazards of water and wind erosion. All seriously disturbed land areas are in need of reclamation and should be stabilized and reclaimed as quickly as possible after disturbance. Disturbance comes from major construction projects such as interstate highway systems, shopping centers, and housing developments, and from surface mining operations for coal, stone, gravel, gold, phosphate, iron, uranium, and clay. Surface mining for coal is responsible for almost one-half of the total land area disturbed in the United States, another one-fourth is from sand and gravel, and the remainder is from mining of other materials and construction. *See* EROSION.

Surface-Mined Areas

The amount of coal recoverable by surface mining is not known since mining technology changes rapidly, and overburden removal is basically a question of economics. The coal required to meet the needs of power production cannot be supplied from deep mining. This means that thousands of acres will be added annually to those already disturbed. In the United States, federal and state laws require that coal exploration and surface coal mining and reclamation operations be conducted in manners that are compatible with the environment and the economic and social needs of each state. The laws also require that reclamation efforts, including but not limited to back filling, grading, top soil replacement, and revegetation of all land disturbed by surface mining activities, shall occur as contemporaneously as practicable with mining operations. Regulations require immediate precautionary measures to be taken to stabilize these severely disturbed areas to avoid serious pollution of the environment.

Surface mining creates many environmental problems. The chemical and physical properties of the resulting spoil is drastically changed and can create a hostile environment for seed germination and subsequent plant growth. However, many of these environmental problems can be overcome or eliminated by proper planning prior to mining and by proper removal and replacement of overburden material during mining operations. Federal and state laws further provide for the control of overland water flow by requiring the construction of temporary diversions as well as permanent diversions to handle peak water flow as needed in order to safeguard surrounding areas from environmental damage. Even with the best mining practices, vegetative cover must be established almost immediately, or the denuded areas will be subject to both wind and water erosion that will pollute surrounding streams with sediment. *See* COAL MINING; MINING; SURFACE MINING; WATER POLLUTION.

In most cases, the spoil material is a conglomerate of rock fragments of various sizes mixed with a small amount of soil. The spoil materials may vary in color from very light to almost black. Because of these color differences alone, temperature and soil moisture may constitute serious problems. Temperature variations created by slope and aspect require use of specific plant species. Mulching is usually essential for germination and seedling establishment. Some spoils are almost sterile and usually require applications of many of the nutrient elements necessary for plant growth. Nitrogen, phosphorus, potassium, calcium, and magnesium deficiencies are major problems on some spoils in the eastern United States, while in other spoils some plant nutrients may be more abundant after mining than before because of weathering of rocks and minerals brought to the surface. Chemical constraints for reclamation in humid areas are low fertility, low pH or high acidity, and toxic concentrations of elements such as aluminum, manganese, and iron. *See* PLANT MINERAL NUTRITION; SOIL CHEMISTRY; WEATHERING PROCESSES.

Arid regions. Under arid conditions of the western United States, excessive amounts of sodium and other elements may constitute serious reclamation problems. High-spodic (heavy, claylike) spoils may require applications of surface soil before good vegetative growth can be accomplished. Under arid and semiarid conditions, lack of plant-available soil water is a factor that ultimately limits plant growth on reclaimed land, just as on unmined land. Practices that increase infiltration, reduce evaporation, and increase potential plant growth generally improve the effectiveness of water conservation and use. Environmental problems of major concern under western arid conditions include salinity levels, exchangeable sodium content, nutrient deficiencies (nitrogen and phosphorus), plant toxicities (magnesium, boron, molybdenum), soil compaction, and steepness of slope.

Alpine and arctic areas. Reclamation and rehabilitation of disturbed land areas in alpine and arctic regions constitute a special problem. These areas are found mainly at high elevations in the western states and the Alaskan arctic tundra area. Examples of land disturbance in arctic and alpine areas involve the Alaskan pipeline and discoveries of large deposits of minerals at high elevations in some of the western states, such as the alpine tundra of the Beartooth Mountain of Montana where chromium and platinum are being surface-mined.

Rehabilitation of these areas is complicated by the short growing season for plants, infertile soils, and the rigorous environments. Some of the greatest problems encountered in reclamation are on areas where total loss of soil has exposed acid-producing pyrites and high concentrations of toxic metals. In some cases, reclamation is aggravated by high winds that remove the exposed soil and sweep away insulating materials, such as snow. *See* TUNDRA.

Topsoil replacement. In the United States the Federal Strip Mine Law requires that topsoil be removed and reapplied on the spoil surface during regrading and reclamation. This practice alone has aided materially in reclamation of surface mine spoil areas throughout the United States. Even when topsoil is reapplied, the surface may contain coarse-textured materials and rock fragments, making it difficult to establish vegetative cover. Many of the eastern mine spoils are derived from sandstone and shales and have a low water-holding capacity. These spoils tend to form crusts and thus create a water-impermeable layer. Practically all of these topsoils have low fertility and thus require extensive fertilization for reclamation and seedling establishment. *See* GROUND-WATER HYDROLOGY.

Placement of overburden. In the eastern United States, many of the surface soils underlined by coal have very low productivity. With proper placement of overburden during mining, use of better regrading techniques, and proper surface management, many surface mine areas can be more productive after mining than before. Some overburden materials contain large amounts of calcium, magnesium, phosphorus, and potassium. If these materials are brought to, or near, the soil surface, weathering may provide a better soil than was there initially.

Use of waste materials. A large number of domestic and industrial waste materials have been tested as amendments on surface mine spoils. These include digested sewage sludges, composted sewage sludge and garbage, tannery waste, bark and fiber mulches, flyash, fluidized-bed combustion waste, and scrubber sludges. Many of these materials contain considerable amounts of essential plant nutrients, but they may also contain some potentially toxic elements. Organic waste materials such as composted sewage and garbage and sewage sludges have been utilized effectively without any detrimental effect. *See* ENVIRONMENTAL ENGINEERING; SOIL.

Wetlands

The term wetlands has been defined in many different ways. In general terms, wetlands are areas that are inundated or saturated by surface or ground water with sufficient frequency and duration to support vegetation typically adapted for life in saturated soil conditions. Wetlands generally include swamps, marshes, bogs and similar areas. In general, wetlands can be classified as either tidal (including saltwater and brackish water, marshes, intertidal flats, and mangrove swamps) or nontidal (occurring inland and including swamps, bogs, ponds, bottomland, hardwood forests, and areas along streams and lakes).

According to the United States Fish and Wildlife Service, over 50% of the wetlands (approximately 11×10^6 acres or 4.4×10^6 hectares) have been destroyed during the last two centuries. Most of the losses were due to drainage of inland wet areas for agricultural purposes. The development of urban areas accounted for about 8% of the total. Wetlands provide a haven for a multitude of fish and wildlife as well as protection against floods and contaminated water. Coastal wetlands are a vital ecological resource, because they provide shelters and spawning grounds for large numbers of commercially valuable fish. In many land reclamation projects, the value of wetlands is recognized as a vital part of the overall picture. Not only do they provide for spawning and nesting areas for marine and wildlife, but they also provide a natural means of recycling waste materials and cleansing polluted and silt-laden water. A number of communities are taking advantage of the ability of wetlands, both natural and constructed, to treat tertiary wastewater and stormwater.

Preservation. Even though many laws and regulations are in effect to protect wetlands, they still face constant threats from real-estate development and agricultural uses. Developers have practiced converting wetland areas by filling with sand, gravel, soil, and other material to give stable substrate for building sites. This procedure tends to significantly alter adjoining areas by restricting both surface- and ground-water flow. Farmers usually reclaim wetlands for crop production by draining and diking. Despite the laws and programs to protect wetlands, demographic trends, especially along the coastal areas, will continue to increase pressure for development of these areas. Nevertheless, laws and regulations are helping through preservation and a procedure known as mitigation. This term, in general, refers to activities that can avoid or minimize damage to wetlands and, in addition, restore, enhance, or create new wetlands. Thus, mitigation may result in no net loss of wetlands. While the process may reduce wetland losses, mitigation is not without controversy.

Sometimes mitigation involves certain destruction of a wetland site in exchange for uncertain gains. Wetlands involve complex ecosystems of soil, water, and plant and animal life that are difficult to recreate. Nevertheless, mitigation can be successful if appropriate guidelines are followed.

Restoration. Restoration of wetlands, rather than mitigation, is preferred by some states, believing that restoration involves less risk than trying to develop new upland sites. The abused sites can be restored as self-sustaining ecosystems by cutting off pollution discharge, removing fill, and introducing native wetland plants. Restoration can be enhanced by bringing bottom layers of soil and muck from existing wetlands and spreading it on the site to be restored.

Tidal wetlands can be reclaimed by construction canals and opening up access waterways to the main body of water. In many areas this may mean removing some fill dirt to allow free movement of water between high and low tides. This step should be followed by introducing plant species such as salt hay and cord grass in areas along the Atlantic and Gulf coast, and glasswort and Pacific cord grass along the Pacific coast. Many communities along coastal shorelines have established coastal wetlands to control erosion.

When reclaiming nontidal wetlands, a greater variety of plant species can be used, but will depend on type of environmental conditions sought. Freshwater emergent marshes can be seeded with species such as fox sedge, soft rushes, cattail, reed canary grass, and purple loose strife after necessary conditions have been established by removing drainage ways and fill dirt.

Actually, reestablishing wetlands can be accomplished by reversing the process by which they were destroyed. To speed up the reclamation process, it will usually be necessary to remove introduced contaminants and fill. Often marshes, bogs, and swamp areas have been used as dumping grounds in place of approved landfills. Removal of most of this trash is necessary, since it may contain materials and chemicals toxic to both plant and animal life. After cleanup has been accomplished, these areas should be monitored annually by making chemical, physical, and biological evaluations of the soil, water, and plant and animal life. If desirable plant species are not present, they should be introduced as soon as possible by seeding or transplanting once a desirable environment has been established; marine and terrestrial wildlife will follow. Many of the bogs near population centers are in need of cleanup and restoration, especially along the east coast and in Michigan and Wisconsin. Plant species normally adapted to such areas include cranberry, pitcher plant, cattail, Venus' flytrap, evergreen shrubs, and stunted trees. *See* ECOSYSTEM. Orus L. Bennett

Bibliography. W. H. Armiger, J. N. Jones, Jr., and O. L. Bennett, *Revegetation of Land Disturbed by Strip Mining in Appalachia*, Agricultural Research Service, ARS-Ne-71, 1976; W. Mitsch, *Wetlands*, 2d ed., 1993; D. Salvensen, *Wetlands: Mitigating and Regulating Development Impacts*, 2d ed., 1994; Society of Agricultural Engineers, *Land Reclamation: Advances in Research and Technology*, 1993; R. W. Tiner, *Wetlands Identification, Delineation, and Classification*, 1995; F. W. Schaller and P. Sutton (eds.), *Reclamation of Drastically Disturbed Lands: A Proceedings of a Symposium*, 1978.

Land-use classes

Categories of land resources that provide information about uses and characteristics. In combination with geographic location systems, measurements of the amount of land in use, the amount available for development, and the changes in use over time, decisions about land use can be accurately made. Concerns develop when a certain land quality is required for a specific purpose, such as food production, human living areas, or environmental protection.

Large changes have occurred since the early 1960s concerning land-use classification and natural resource management. Previous concerns were allied to very specific problems such as soil erosion, land availability, desertification, or forest consumption. Technologies and theories now view land-use classification in a larger context and relate it to human use of land-based resources within environmentally acceptable guidelines. Major technological improvements that contributed to these changes include high-resolution films and improved cameras, film processing, lenses, and enlargers. Airplanes, helicopters, and satellites provide platforms for highly developed scanners, return-beam vidicon cameras, radar, lasers, and multispectral retrieval systems. Computerized instrumentation allows merging of numerous sources of data for selected areas of the Earth's surface, including continental or global coverage. Thus, completely new approaches have become feasible, such as the *Land Use Classification System for Use with Remote Sensor Data* prepared by the U.S. Department of the Interior, and previous sources of land-use information have been maintained or enhanced. *See* REMOTE SENSING.

Land is readily classified according to a group of natural characteristics that occur in association. H. E. Conklin developed land classification systems based on soil, climate, topography, and location. Others have preferred to add criteria based on bedrock geology, vegetative cover, subsurface characteristics, landforms, and hydrologic features.

Inventories. Land-use inventories have been encouraged by professional organizations, such as the International Geographical Union. Historically, land-use classification has been based on the concepts of cadastral surveys, which originally included land classification criteria based on quality, quantity, location, and ownership of the land resource. The objective was to provide a basis for equitable taxation. The current objective of land-use classification, however, also includes information for resource management.

Inventories such as the U.S. Department of Agriculture System of Land Capability Classification have centered on the suitability of soils and their limitations for field crop production. The limitations are broadly identified as few, moderate, severe, very severe, impractical to remove, generally too severe for cultivation, very severe limitations for cultivation, or too severe for commercial crop production. Similar major land resource inventories have been developed in Canada, Australia, United States, Netherlands, Denmark, Sweden, England, Switzerland, and other European countries.

Two technologically advanced major inventories, the New York State Land Use and Natural Resource Inventory (LUNR) and the Canada Land Inventory, were developed in the late 1960s for very large areas. The LUNR system provided geographically referenced land-use information on 155 items of information derived primarily from aerial photographs. These can be grouped under the 11 major use headings of agriculture, forest land, water resources, wetlands, residential, commercial and industrial, outdoor recreation, extractive industry, public and semipublic, transportation, and nonproductive uses. Each group is subdivided into 2 to 25 subunits, offering a very detailed classification system for uses of major importance, such as agriculture or residential. Its products included maps, statistical data, computer graphics, and computer printout. The LUNR program sparked activity in land-use classification throughout the United States, and in a few years 32 states developed similar land-use classification systems. The Canada Land Inventory functioned on different concepts, providing highly refined, geographically located in formation on a relatively small number of land-use classification items with computer retrieval capabilities.

Principles. The ideal land-use classification system would provide appropriate identification for all land uses with mutually exclusive units and boundaries clearly defined and easily understood. A thorough description of each unit would describe what the unit contained and why it was placed in its particular classification mode. But natural resources, vegetation, and land patterns influenced by civilization occur in a continuum with transition zones, instead of clearly defined boundaries, between land uses. As a result, generalization of the classification units is required. Because land-based resources are geographically fixed, impurities and mixed units must be included to be comprehensive. The purely theoretical approach may look at vegetative patterns, natural successional systems, and modification of human-impact consequences to maintain the purest possible scientific sequence.

The more common approach is to respond to the request of the sponsor. This usually means forfeiting some of the desired theoretical concepts to maintain cost effectiveness. Steps in this process include: (1) identification of the sponsor's information needs; (2) selection of sources of information; (3) development of a classification system that is comprehensive and provides unique descriptions and discrete assignment; (4) preparation of a geographic referencing system; and (5) selection of appropriate information-retrieval processes.

The classification system should be able to stand alone, independent of other component parts of the inventory project. It should not be dependent on unique talent, special instrumentation, or any one source of information. Classification and inventory decisions do not require computer or other forms or instrument capability; the use of computers is primarily to store, manipulate, and retrieve information.

Technological development. Development of available technologies has enabled many changes to occur in analytical methodologies. There has been rapid proliferation of methods for analysis of the capability of ecological systems to accommodate particular human land-use needs. These have been accompanied by advances in predictive evaluation of regional landscapes, and have thus carried the process into areas of systems modeling.

Methods for land-use capability analysis have progressed from trial-and-error techniques, which were among the early clearly identifiable regional planning methods of the twentieth century, to inventory methods, characterized by systematic inventories of land uses that combine information from natural sciences with engineering classification of soils and aerial photographic interpretations of cultural land-use patterns. Hand-drawn overlay methods came next to address the complex decisions to be handled at a regional level of planning. Finer distinctions between land uses and resources, and better elucidation of relationships between different types of land use were accomplished by hand-drawn overlay methods than by inventory mapping. Subsequently, data classification systems emerged, as information theory and systems analysis were employed as the organizing elements in data classifications. Computer-based planning methods, including computer-aided assessment procedures, later enabled extensive use of quantitative weighting systems and the sophisticated techniques of derived weights and values.

In the early 1900s in North America, soil and physical resource data were often incomplete and were published in literary form and also at different scales which made comparisons of data very difficult. Later the development of inventory methods was made possible by the considerable increase in knowledge about soils, geology, hydrology, and ecology. Innovations in remote sensing, refinements in soil taxonomy, and enhanced cartographic representations added to the greater availability of natural resource and land-use data. With hand-drawn overlay methods, maps of data developed at a common scale and overlaid are employed in various combinations to create a new level of information: composite maps.

Geographic information systems help serve the interest in finer distinctions between land uses and resources, and the attempts at resolution of conflicts by analysis of relationships between categories of land use. These may be found in the United States' state and regional systems, as well as those in western Europe, Canada, and some developing countries. In the computer era, faster data sources have become more comprehensive, leading to more complex site analysis and land-use planning. Highly advanced landscape planning methods have resulted, which include the use of computer-displayed three-dimensional forms.

In the Netherlands, basic material on the natural environment has been collected over a long period of time. Since 1970, such data has been made suitable for use in behalf of national physical planning policy. A national environmental survey begun in 1972 had as its purpose the elaboration of theoretically demonstrated relationships between community and natural environment for practical situations. This survey and a General Ecological Model (GEM) serve to illustrate the Netherlands' commitment to use of the science of landscape ecology in planning.

A. Lieberman; Ernest E. Hardy

Bibliography. J. R. Anderson et al., *A Land Use and Land Cover Classification System for Use with Remote Sensor Data*, U.S. Government Printing Office, Survey Prof. Pap. 964, 1976; J. G. Fabos, *Planning the Total Landscape: A Guide to Intelligent Land Use*, 1979; E. E. Hardy, *The Design of Natural Resource Inventories*, Resource Information Laboratory, New York State Cooperative Extension, 1979; F. J. Marschner, *Boundaries and Records*, Agricultural Research Service, USDA, 1960; Z. Naveh and A. S. Lieberman, *Landscape Ecology*, 1985.

Land-use planning

The long-term development or conservation of an area and the establishment of a relationship between local objectives and regional goals. Land-use planning is often guided by laws and regulations. The legal basis of land-use regulation in the United States is the police power of a city (or county) to protect the health, safety, and welfare of its residents. The major instrument for current land-use planning is the establishment of zones that divide an area into districts which are subject to specified regulations. Although land-use planning is sometimes done by private property owners, the term usually refers to permitting by government agencies. Land-use planning is conducted at a variety of scales, from plans by local city governments to regulations by federal agencies. The United States has never developed a national land-use plan because land use is considered a local concern.

History. Ancient cities were often designed on rectangular grid patterns and surrounded by protective walls. Modern land-use planning began in the Italian Renaissance during the fifteenth century. Antonio Filarete planned an ideal city, Sforzinda, star-shaped with a radial road pattern. This was an early attempt to organize the land use of an entire city. Early visionaries focused on the form and shape of cities. For example, the inventor Ebenezer Howard introduced the English concept of a "garden city" in the 1890s. In his design, 30,000 people were to live on 1000 acres of land, separated from the city by a greenbelt that was 5000 acres in extent. Homes were laid out in a circular pattern around the city. The design was motivated to overcome the problems of industrial cities. Howard's garden city later gave rise to planned unit developments (PUD), a common tool for developers in suburban areas. Until the early twentieth century, land-use planning was mostly a physical solution to land-use issues in urban areas.

American land-use planning has an erratic history that draws on a broad variety of disciplines, such as architecture, civic design, parks and open spaces, fire prevention, public health, and housing. An early American planner, Frederick Law Olmsted (1822–1903), was important in the development of land-use planning. His ideas went beyond traditional bounds of shape and form and focused on the functions of different pieces of land, concepts that were a major foundation block in the evolution of zoning. In his plans for the metropolitan area of Boston, he developed some of the first major linear open spaces around the city. He recognized the natural landscape not just as a cultural and esthetic resource but also as an important component in public health. He argued that the closely built walled cities of medieval Europe were responsible for much sickness at that time. Soon after, Charles Eliot planned public open space for Boston around the idea of high-density housing near trolley lines and extensive open spaces in the interstices, along with numerous small squares and parks in the midst of dense populations.

The first comprehensive zoning ordinance in the United States was written by New York City in 1916. It stipulated the regulation of height, area, and use of buildings for public health and safety in various districts. The essence of this ordinance was the designation of different provisions in different areas. In *Village of Euclid, Ohio v. Ambler Realty Co.* in 1926, the U.S. Supreme Court ruled that zoning laws were a valid extension of nuisance laws. The ruling implied that nuisance can be related to the location of a problem rather than just the nature of a problem, such as a pig in a parlor instead of a pig in a pen.

Local-level planning. Most land-use planning occurs at the local level in the United States. American citizens value local autonomy and flexibility in land use. According to one expert, land-use planning in the United States is not a system but 40,000 city and county governments independently administering zoning. Most planning is decided by boards or commissions at the city level, often composed of volunteers or elected officials who are assisted by professional planners. The general public contributes to planning through public hearings and written comments submitted to the planning bodies.

Each local government establishes a general plan which acts as the basic planning document. The general plan indicates the long-term goals for the physical development of the city and any land outside the city's borders that the local government judges to relate to its planning. These plans vary by city and by state. In California, for example, the general plan must have seven elements: a land-use element, a circulation element, a housing element, a conservation element, an open space element, a noise element, and a safety element. The land-use element includes the distribution and intensity of land uses and population density measures. Specific plans are then written to systematically implement the general plan in specific areas. These plans often follow a general process that includes formulation of goals, assembly of data, analysis of data, presentation of results, and implementation.

A major part of local planning is zoning, the division of areas into districts. Zones cover most potential uses, such as residential, commercial, light industry, heavy industry, open space, or transportation infrastructure (such as rail lines or highways). Detailed regulations guide how each zone can be used. For example, residential zoning regulations may cover the height of houses, the distance the houses are placed from the street, or the colors houses can be painted. Open space zoning may include which spaces are designated as picnic areas, as playing fields, as hiking trails, or as strictly wilderness. The placing of different zones relative to each other is increasingly taken into consideration. Zoning has often been used by wealthy communities to promote low-density housing. The use of zoning to exclude certain groups of people has long been criticized.

As a result of pressures from rapid growth, some cities have begun to write growth management plans that limit the pace of growth. Comprehensive city plans aimed to limit the pace of growth have been accepted by the courts. *See* LANDSCAPE ARCHITECTURE.

State-level planning. Very few plans have been undertaken at a statewide scale. In the 1970s, some states developed regional plans to address urgent growth pressures. Each state plan differs by the needs and philosophy of the state. For example, Hawaii introduced statewide zoning to address rapid urbanization of valuable agricultural land; Vermont set up a system of citizen district commissions to administer the state's planning system; and Oregon, in response to large losses of public access to the coast, developed a comprehensive system of local planning in which local plans must conform to specific state goals. The state plans represent a balance of regional structures that address widespread growth with local powers that keep specific decision-making at the local level.

National-level policies. In the United States, very few land-use laws or policies originate at the national level. One of the earliest exceptions was the 1785 Land Ordinance which initiated rectangular surveys in areas west of the Appalachian Mountains. The surveys established 6 mi × 6 mi townships, each with 36 sections of 640 acres (1 mi × 1 mi). The surveys were instrumental in the rapid and orderly land utilization of vast areas. Based on these surveys, the Railway Land Grant gave away over 150 million acres of land to the railroad between 1850 and 1871, mostly in 640-acre sections. These early acts had a major impact on the long-term land use of the United States, and its effects are still emerging in many issues today. For example, the railroad was given every other section in parts of the West, with the remainder staying in public hands. In forested areas, the resulting pattern is that of a checkerboard across vast areas because the railroad's property was sold and logged by private timber companies, leaving a pattern of clearcuts and mature forest.

Environmental regulations are among the few national-level policies that have direct implications for land-use planning. Four of the major types of environmental laws that impact land-use planning are wetland laws, clean-air laws, clean-water laws, and laws for the protection of endangered species. Wetlands are among the most productive ecosystems, with important roles in flood control, nutrient and pollution removal, and fish and wildlife habitat. Wetlands are defined as those areas inundated or saturated by water at a frequency and duration sufficient to support a prevalence of vegetation typically adapted for life in saturated soil conditions. The U.S. Army Corps of Engineers (ACOE) is responsible for implementing wetland regulations. The ACOE permits development in wetland areas and decides if sites restored to mitigate for wetland losses have been suitably restored. In addition to enforcing wetland laws, the ACOE enforces many regulations that resulted from the Clean Water Act of 1972. The object of this act is to restore and maintain the biological, physical, and chemical integrity of the nation's waters. The Environmental Protection Agency (EPA) oversees standards for clean air that were first established by the Clean Air Act. Although laws to maintain clean air have less direct impact on land use, they have many indirect impacts. For example, the clean-air laws impact what kind of industry can be developed in different places.

The Endangered Species Act of 1974 significantly impacted land-use decisions. It mandates protection of all habitat for endangered and threatened animals and protection of all habitat on public land for endangered and threatened plants. The act has four main components: (1) provisions to list species as threatened or endangered, (2) requirements for federal agencies to consult with the U.S. Fish and Wildlife Service (USFWS) or the National Marine Fisheries Service (NMFS) on federal projects, (3) prohibitions against taking, and (4) provisions for permits to allow incidental take under special circumstances. Species are listed as endangered or threatened if their continued existence is determined to be in jeopardy. Causes for a jeopardy ruling can include threat of habitat destruction, threat from overuse, threat from predation or disease, inadequacy of existing regulation, or other factors that threaten a species' existence. *See* ENDANGERED SPECIES.

In 1982 the act was amended to allow incidental taking of threatened or endangered species. To qualify, the taking must be "incidental to, and not the purpose of, the carrying out of an otherwise lawful activity." The amendment applies to private landowners and nonfederal entities, provided they develop agreements known as habitat conservation plans. These plans agree to minimize and mitigate any taking. These agreements are quickly becoming a prominent mechanism to address threatened and endangered species on private land and had a major impact on land use in the 1990s. Over 200 were approved during the 1990s, spanning scales from private homeowners with less than an acre of land to agencies of state governments that have received permits to use thousands of acres of timberland.

From city to landscape. Land-use planning, in large part, has focused on urban planning. Most texts on the subject dwell on urban plans, zoning (especially residential zoning), and laws implementing zoning at a local level. Increasingly, land-use planning is done at larger scales and involves multiple issues. Awareness of environmental concerns, coupled with the wide availability of technical tools that include digital maps at all scales, has led to new approaches to land-use planning. These approaches often use ideas from landscape ecology in the planning process.

Landscape ecology embraces large, heterogeneous areas. It began in the 1950s as a field of study that combined natural history and physical patterns of large areas with aerial photographs. Four general areas of landscape ecology have direct implications for land-use planning: patches; edges, boundaries, and fragmentation; buffer zones; and corridors and connectivity. *See* ECOLOGY, APPLIED; LANDSCAPE ECOLOGY.

Three features of patches are important in land-use planning: their size, location, and abundance. In general, larger patches are better than smaller patches in maintaining wildlife and vegetation. Similarly, a greater number of intact natural patches is helpful. Development of patches of natural areas may result in loss of rare species and fewer kinds of habitats. It is also better to keep distances between patches short. For the land-use planner, this impacts decisions about which areas to develop. For example, land-use plans for a housing development might consider which areas to develop and which natural areas to leave intact in terms of where natural habitats remain and the size of individual habitat patches, rather than considering only the total acreage of native habitat in the vicinity.

Patch edges often have a different kind of habitat than the middle of a patch. Often this boundary has a negative impact on species that live in the middle of a patch. One reason that larger patches are better for wildlife is that a smaller percentage of the habitat comes in contact with developed areas. Similarly, patches that are more circular in shape maintain more interior habitat than those that have a less regular shape. For this reason, when habitats are fragmented by development into odd shapes and thin strips, little interior habitat area may remain. Also, abrupt edges may have a more negative influence on wildlife than less distinct edges. For example, an abrupt edge may reduce soil erosion and decrease habitat diversity. Thus, land-use plans that impact wildlife areas should consider the shape of patches and boundaries between natural areas and developed areas when designing specific plans.

In natural areas with a high sensitivity to development, a buffer zone provides a protected area between a sensitive area and a developed area. Buffer zones are areas that land-use planners can designate for low-intensity activities, such as parks and recreation.

Many ecological processes are influenced by the connections between patches of natural habitat. Animals may move freely in corridors of native habitat or may be inhibited by large barriers such as interstate highways. In some terrestrial cases, stepping stones of similar habitat may provide enough territory to provide connectivity. Similarly, a stream and its banks must be continuous, not broken, for maintenance of an aquatic community. The connection between sensitive wildlife patches is particularly important in land-use planning because planners must consider how to leave corridors of open space intact and unbroken by development. Cheryl Schultz

Bibliography. J. B. Cullingworth, *The Political Culture of Planning: American Land Use Planning in Comparative Perspective*, 1993; D. J. Curtin, Jr., *California Land Use and Planning Law*, 1994; W. E. Dramstad, J. D. Olson, and R. T. T. Forman, *Landscape Ecology Principles in Landscape Architecture and Land-Use Planning*, 1996; J. G. Fabos, *Land-Use Planning: From Global to Local Challenge*, 1985.

Landing gear

That portion of an aircraft consisting of the wheels, tires, brakes, energy absorption mechanism, and drag brace. The landing gear is also referred to as the aircraft undercarriage. Additional components attached to and functioning with the landing gear may include retracting mechanisms, steering devices, shimmy dampers, and door panels.

The landing gear supports the aircraft on the ground and provides a means of moving it. It also serves as the primary means of absorbing the large amounts of energy developed in the transition from flight to ground roll during a landing approach. The brakes, normally located in the main wheels, are used to retard the forward motion of the aircraft on the ground and may provide some control in the steering of the aircraft. In most modern aircraft the landing gear is designed to retract into the aircraft so that it is out of the airstream and drag is thus reduced.

Types. Early aircraft and many small aircraft use a tail-wheel (or skid) in a conventional, or tail-dragger arrangement, in which the main landing gear is located ahead or forward of the center of gravity of the aircraft (**Fig. 1***a*). The popular arrangement on modern aircraft is a tricycle landing gear, with the main gear located behind or aft of the center of gravity, and a nose gear located forward which carries about 20% of the static weight of the aircraft (Fig. 1*b*). Other landing gear arrangements include four wheels, with two wheels on either side of the aircraft in a tandem arrangement; tandem wheel configurations, with the gear mounted on the centerline of the aircraft; single-wheel gear, such as that used for sailplanes; and skid-type gear, often used on helicopters. Skis or floats may be used in lieu of wheels. *See* GLIDER; HELICOPTER; SEAPLANE.

Large aircraft such as the wide-body commercial aircraft and military aircraft like the C-5A employ multiple-wheeled bogies to support their huge weight and, in the case of the C-5A, to provide soft terrain landing and takeoff capability. The bogie landing gear arrangement on the XB-70 aircraft (**Fig. 2**) employs a unique braking system using a small sensing wheel, located between the main tires, to sense impending skid due to braking by measuring the ground speed compared with the wheel speed of the main tires. Several aircraft also employ outrigger landing gear on the wing tips for roll stability. *See* MILITARY AIRCRAFT.

Fig. 1. Landing gear arrangements. (*a*) Tail-wheel arrangement. (*b*) Tricycle arrangement.

Design considerations. Landing gear design parameters are weight, vertical velocity of the aircraft at gear contact, landing approach speed, braking distance, ground clearance, lift angle requirement, turnover angle (related to the distance between the two main gears and also the distance between the nose gear and main gear), the condition of the landing surface, and whether the landing is on the ground or an aircraft carrier. Once the basic geometric relationships have been defined, the most important design consideration is determining the amount of energy that must be absorbed by the landing gear to prevent overloading the aircraft structure. The ideal situation is to match the loads induced into the structure to those induced by the various flight conditions.

Oleo strut. The most accepted method of absorbing the energy due to landing is an air-oil strut called an oleo. The basic components are an outer cylinder which contains the air–oil mixture and an inner piston that compresses the oil through an orifice. The flow of oil through the orifice is metered by a variable-diameter pin that passes through the orifice as the gear strokes. The flow of oil in effect varies the stiffness of the compression of the gear. Because of the flexibility of the aircraft and gear and the smoothness and strength of the landing surface

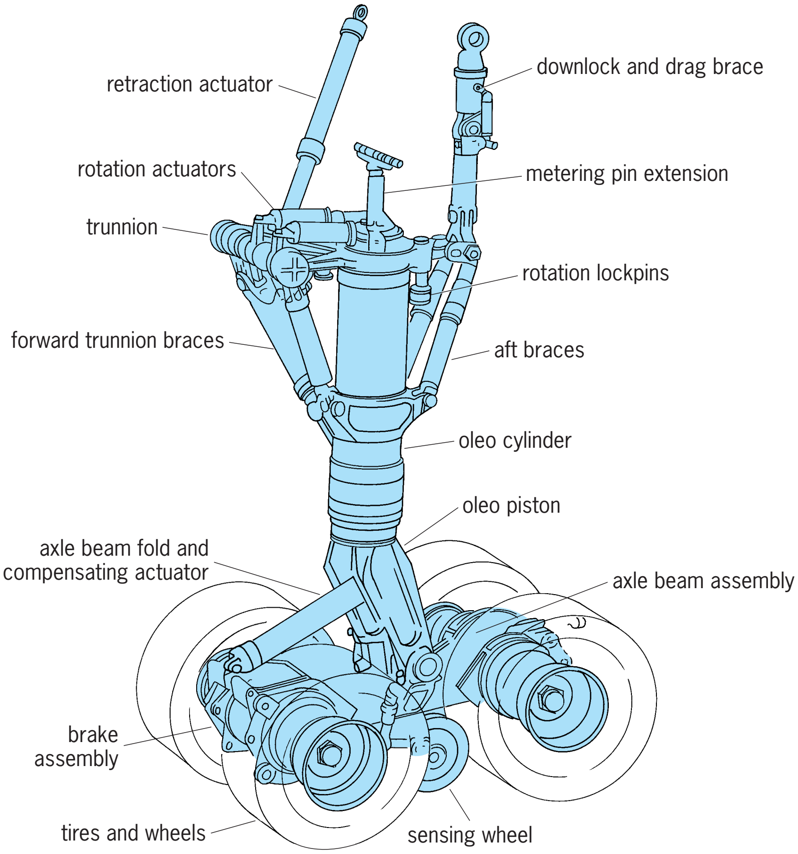


Fig. 2. Main-wheel bogie for the XB-70A aircraft. (*After S. Pace, North American Valkyrie XB-70A, Aero Series vol. 30, Tab Books, 1984*)

metering is necessary to establish the most efficient load–stroke relationship. Thus the overall stroke of the gear is minimized, and the complexity, flexibility, and strength requirements are optimized.

Special requirements. Naval aircraft must be able to land aboard and take off from aircraft carriers. The sink speeds can be as high as 24 ft/s (7.3 m/s), compared with 10 ft/s (3.0 m/s) for Air Force aircraft and 3 ft/s (0.9 m/s) for most commercial aircraft. The high energy generated requires long gear stroking and special metering characteristics. A secondary metering orifice is often required for arrested landings, particularly free-flight engagements. A free-flight engagement occurs when the arresting cable is engaged prior to the aircraft wheels touching down. Very high energy is imparted to the nose gear. The aircraft comes to a stop within approximately 310 ft (100 m). This results in high longitudinal load factors (4–5 *g*) on the aircraft.

The nose gear is specially designed for catapult launches. The design includes a retractable tow bar which drops over the catapult shuttle into a notch. The tow bar is attached to the nose gear strut and transmits the launch force from the catapult to the aircraft through the trunnion and a drag brace. The tow bar also serves to align the aircraft as it approaches the catapult by engaging a track on the carrier deck. A holdback bar trails behind the nose wheel and drops into a slot in the deck. A holdback coupon with a predetermined breaking load based on aircraft weight and catapult characteristics connects the holdback bar to the holdback fitting. This allows the launch load to build, imparting a step acceleration to the aircraft. The longitudinal acceleration during catapulting can exceed 5 *g* with corresponding loads in the launch bar over 250,000 lbf (10^6 newtons). The length of the power stroke of the catapult is approximately 250 ft (80 m), with flying speed being attained in this distance. Dual nose wheel tires are required so that the catapult shuttle can be straddled and maintain symmetric load on the trunnion. *See* NAVAL SURFACE SHIP.

Aircraft that operate from unprepared fields also have special requirements. Large, low-pressure tires are needed to provide a high flotation capability. Since the surface may be rough and contain obstacles such as rocks and logs, the gear is likely to employ additional energy absorption mechanisms. A trailing-arm (articulated) gear is used on the OV-10A aircraft, allowing it to land in soft fields with high sink rates. Active landing gear designs employ computer technology within the gear to sense the rate of onset of load and actively meter the flow of oil to compensate and maintain an optimum load level. *See* AIRPLANE.

Ralph R. Deakins

Bibliography. K. Connes, *Know Your Airplane*, 1986; N. S. Currey, *Aircraft Landing Gear Design: Principles and Practices*, 1988; W. G. Holder, *The B–1 Bomber*, Aero Series vol. 35, 1988; S. Pace, *North American Valkyrie XB70A*, Aero Series vol. 30, 2d ed., 1989; H. Seo, *Boeing 747*, 1984; D. B. Thurston, *Design for Flying*, 2d ed., 1994.

Landscape architecture

The art and profession of designing and planning landscapes. Landscape architects are concerned with improving the ways in which people interact with the landscape, as well as with reducing the negative impacts that human use has upon sensitive landscapes. The history of landscape architecture reaches from the gardens and outdoor environments of ancient civilizations to a broad base in environmental design in the twentieth century. Today, landscape architects are involved in such diverse areas as landscape and urban design, community and regional planning, interior and exterior garden design, appropriate and high technology, agricultural and rural land-use planning, parks and recreation, historic site and natural area preservation, landscape restoration and management, research and academic programs, energy and water conservation, and environmental planning. Landscape architects find employment with private landscape architecture, engineering, and planning firms; public agencies, including city, county, and state design, planning, park, and transportation department; federal agencies, including the U.S. Forest Service, Corps of Engineers, National Park Service, Bureau of Land Management, and Fish and Wildlife Service; and universities and colleges.

The American Society of Landscape Architects

(ASLA) was established in 1899 as the primary informational and lobbying organization for the profession. Over 80% of the states in America have licensure for landscape architects through state consumer, health, and safety regulation boards. There are over 40 accredited college programs in landscape architecture throughout the United States which provide educational training, enabling students to take licensing exams without lengthy apprenticeships. Degrees in landscape architecture are offered at both the bachelor's and master's levels. Many countries have universities with degree programs in landscape architecture, and abundant international employment opportunities exist. There are landscape architects in practice in every country of the world.

Frederick Law Olmsted, the nineteenth-century designer of Central Park in New York City, Yosemite National Park in California, and many gardens and parks throughout the country, is generally considered to be the father of landscape architecture in America. When Olmsted is added to the long list of historical figures associated with the profession, including the eighteenth-century British landscape architect Humphery Repton and the seventeenth-century French landscape architect Andre Le Notre, the form and historical intent of the designed and planned landscape throughout the world can be better understood.

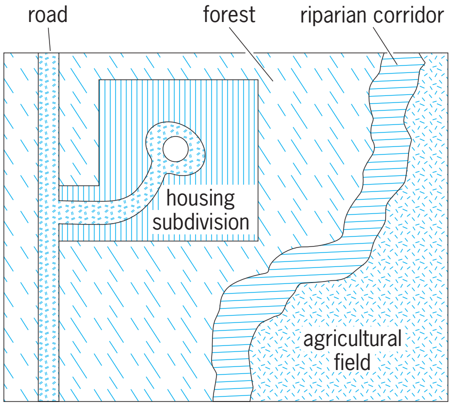
The skills required for landscape architecture are concerned with creativity, ecology, social science, horticulture, engineering, and graphics. Landscape architects are generalists in that their educational and professional experience is very broad. Many environmental and cultural factors affect landscape design and planning, and landscape architects have to know how these factors relate. Design process is the main area of specialization for landscape architects, and decision-making related to design process is the fundamental reason why landscape architects are employed. The ability to utilize the information of specialists and to synthesize the information into a usable format remains the essence of the profession. *See* CIVIL ENGINEERING; ENVIRONMENTAL ENGINEERING; FORESTRY, URBAN; LAND-USE PLANNING.

Kerry J. Dawson

Bibliography. S. Cantor, *Landscape Architecture*, 1992; T. Higuchi, *The Visual and Spatial Structure of Landscapes,* 1988; J. E. Ingels, *Landscaping Principles and Practices*, 5th ed., 1997; T. Turner, *Landscape Planning,* 1987.

Landscape ecology

The study of the distribution and abundance of elements within landscapes, the origins of these elements, and their impacts on organisms and processes. A landscape may be thought of as a heterogeneous assemblage or mosaic of internally uniform elements or patches. Typically, landscape ecologists view the world as one would see it from an airplane (see **illus.**). At this scale, examples of patches include remnant blocks of forest, agricultural fields, and housing subdivisions. Many fields of study focus on large-scale patterns. For example, biogeographers, land-use planners, hydrologists, and ecosystem ecologists are concerned with patterns and processes at large scale. However, landscape ecologists bridge these disciplines in order to understand the interplay between the natural and human factors that influence the development of landscapes, and the impacts of landscape patterns on humans, other organisms, and the flows of materials and energy among patches. Much of landscape ecology is founded on the notion that many observations, such as the persistence of a small mammal population within a forest patch, may be fully understood only by accounting for regional as well as local factors.



Representative landscape as viewed from above.

Although landscape ecology has become a prominent discipline only recently, it originated over 50 years ago in Europe. Important studies in landscape ecology have shown how the sizes of forest patches have declined in England over the last several centuries, that the fires in Yellowstone National Park in 1988 were probably not unusually intense, and that the dispersal movements of owls can be used to develop better methods of quantifying the distribution of habitat within a landscape. One way of understanding the field of landscape ecology is to divide it into three parts: documenting landscape pattern, understanding the development of landscape pattern, and examining the consequences of landscape pattern.

Documenting landscape patterns. A primary task of landscape ecologists is to describe and measure the attributes of landscapes. Landscape ecologists use a wide variety of measures and often employ high-technology tools such as remotely sensed information from Earth-orbiting satellites to obtain the raw data used when documenting landscape patterns. These large spatial datasets are often analyzed utilizing geographic information systems (GIS) on high-speed computers.

Landscapes are defined in terms of patch types. Deciding how to define patch types is a critical step in any landscape ecological study. Will all vegetation be represented as a single patch type, or will forests be differentiated from grasslands? Will forests be divided into coniferous (such as pines) and deciduous (such as oaks)? Will forests of different ages be considered as different patch types? The answers to such questions have an overriding effect on the landscape pattern observed and must be closely linked to the researcher's goal. Once patch types have been determined, frequently used measures of landscapes include the number of patches, their size, and their isolation from other patches. Landscape ecologists also have used a number of measures that reflect the shape of patches. The ratio of patch perimeter to patch area increases as patches proceed from globular to sinuous in their shape. The fractal dimension of a landscape is a measure that can quantify the tendency of very complex patches to fill the space within a landscape.

Development of landscapes. Often, landscape attributes have been quantified as part of an effort to understand the factors that led to the development of the landscape pattern. These factors include a combination of human and nonhuman agents that interact to give landscapes their shape. The geology of a region, including the topography and soils along with the regional climate, is strongly linked to the distribution of surface water and the types of vegetation that can exist on a site. These factors influence the pattern of human settlement and the array of past and present uses of land and water. Thus, fertile lowlands once dominated by floodplain forests have often become agricultural bottomlands. Rocky uplands once sparsely covered with drought-resistant plants have often been used for livestock grazing or the extraction of mineral resources.

In many cases, the role of ancient events can still have strong signatures on the shape of landscapes. Throughout the northern regions of Europe, Asia, and North America, glaciers have repeatedly carved the landscape, eradicated vegetation, and forced plants and animals to reinvade from southern refuges. One consequence of such ice ages has been the prevalence of fresh-water lakes and ponds in glaciated regions.

Acknowledging the important role of geology and climate, there can be little doubt that the landscapes of the world have been dramatically and rapidly transformed in recent centuries. A recurrent theme in landscape ecology is the role of humans in shaping the development of the landscape pattern. One prevalent effect is habitat fragmentation, which arises because humans tend to reduce the size and increase the isolation among patches of native habitat. The clearing of land for human settlement, timber, or agriculture tends to replace irregular, contiguous mosaics of native vegetation with geometric fragments surrounded by a matrix of human-dominated patch types. Much of the attention devoted to habitat fragmentation focuses on the initial clearing of land, such as that occurring in many of the world's tropical forests. However, habitat fragmentation can continue for long periods after this initial transition.

Consequences of landscape patterns. Just as many factors act together to lead to the pattern of patches on a landscape, these patterns can have direct effects on many different processes. Human modification of landscapes results in shapes of landscapes which constrain future human uses. The structure and arrangement of patches also can affect the physical movement of materials such as nutrients or pollutants and the fate of populations of plants and animals. Many of these impacts can be traced to two factors, the role of patch edges and the connectedness among patches.

The boundary between two patches can be important by either hindering or promoting movement. In particular, patch boundaries often act as filters or barriers to the transport of biological and physical elements. As an example, leaving buffer strips of native vegetation along stream courses during logging activities can greatly reduce the amount of sediment and nutrients that reach the stream from the logged area.

The part of a patch near a boundary, or its edge, is exposed to materials and organisms that cross that boundary. Edge effects can result when forests are logged and there is a flux of light and wind into areas formerly located in the interior of a forest. In this example, edges can be a less suitable habitat for plants and animals not able to cope with drier, high-light conditions. When habitats are fragmented, patches eventually can become so small that they are all edge. When this happens, forest interior dwellers may become extinct.

The connectivity of landscapes is directly related to the influence of patch boundaries on the flows among patches. Many species that live in a particular patch type are not easily able to move through other patch types. When patch boundaries act as barriers to movement, they can have pronounced effects on the dynamics of populations within and among patches. In the extreme, low connectivity can result in regional extinction even when a suitable habitat remains. This can occur if populations depend on dispersal from neighboring populations. Habitat fragmentation tends to make it harder for animals to successfully disperse. Distances get longer and the matrix between patches is increasingly unsuitable. Eventually, when a population becomes extinct within a patch, there is no way for a colonist to reach the vacant habitat and reestablish the population. This process is repeated until all of the populations within a region disappear. Landscape ecologists have promoted the use of corridors of native habitat between patches to preserve connectivity despite the fragmentation of a landscape.

An overriding theme in landscape ecology is the role of context. The growth rate of a tree or the likelihood that the tree will survive the winter may depend on the size of the forest patch in which it lives and its distance from the edge of the forest. The presence of a salamander species within the same forest patch may depend on the distance to the nearest neighboring forest patch and whether the intervening space is covered with an abandoned agricultural field or a parking lot. Because landscape ecology focuses on large scales and readily incorporates the role of both humans and natural factors, it has become an important part of conservation efforts: it forces planners to carefully consider their actions across an entire landscape. *See* LAND-USE PLANNING.

David Skelly

Bibliography. R. T. T. Forman, *Land Mosaics: The Ecology of Landscapes and Regions*, Cambridge University Press, Cambridge, 1995; L. Hansson, L. Fahrig, and G. Merriam (eds.), *Mosaic Landscapes and Ecological Processes*, Chapman and Hall, London, 1995.

Landslide

The perceptible downward sliding, falling, or flowing of masses of soil, rock, and debris (mixtures of soil and weathered rock fragments). Landslides range in size from a few cubic yards to over 3.5×10^9 ft^3 (1×10^9 m^3), their velocities range from an inch or so per day to over 300 ft/s (100 m/s), and their displacements may be about an inch to several miles. *See* MASS WASTING.

Types. As with other natural phenomena, difficulties arise in classifying landslides because of the complexity of the processes involved and the limited understanding of these processes. The U.S. Highway Research Board classification is widely accepted, and divides landsliding of rock, soil, and debris, on the basis of the types of movement, into falls, slides, and flows (**Fig. 1**). Other classifications consider flows, along with creep and other kinds of landslides, as general forms of mass wasting.

Falls. This type of landsliding occurs when soil or rock masses free-fall through air. Falls are usually the result of collapse of cliff overhangs which result from undercutting (Fig. 1*a*) by rivers or simply from differential erosion. In temperate and cold regions, ice wedging and roots open cracks parallel to cliff faces, inducing rock slabs, sometimes over 300 ft (100 m) high, to topple and fall (Fig. 1*b*). If these rockfall events are relatively small and piecemeal, the result is an apron of broken rock at the foot of the cliffs known as talus. Talus slopes characteristically assume angles with the horizontal between 30 and 40°.

Slides. Landslides of this type invariably involve shear displacement or failure along one or more narrow zones or planes. Internal deformation of the sliding mass after initial failure depends on the kinetic energy of the moving mass (size and velocity), the distance traveled, and the internal strength of the mass.

Some slides deform very little and move for relatively small distances in comparison with their total size. A slump is a type of these relatively undeformed slides in which the sliding surface or surfaces are spoon-shaped (Fig. 1*c*); in a glide the surface of sliding is planar (Fig. 1*d*). In slumps, the shear plane occurs through intact material; hence, slumps are more common in soil than in rock. In glides, the sliding surfaces are preexisting fractures or weakness planes such as joints, bedding, fault, and foliation planes. Glides are usually associated with rock masses.

Masses of soil, debris, and rock that undergo large deformation during displacement create soil slides, debris slides, and rockslides. Rockslides (Fig. 1*e*) usually involve large (0.35–35×10^6 ft or 0.01–1×10^6 m^3 or more) and extremely large (0.35–35×10^9 ft or 0.01–1×10^9 m^3 or more) rock masses moving for relatively large distances at very high velocities. The most devastating landslides in history

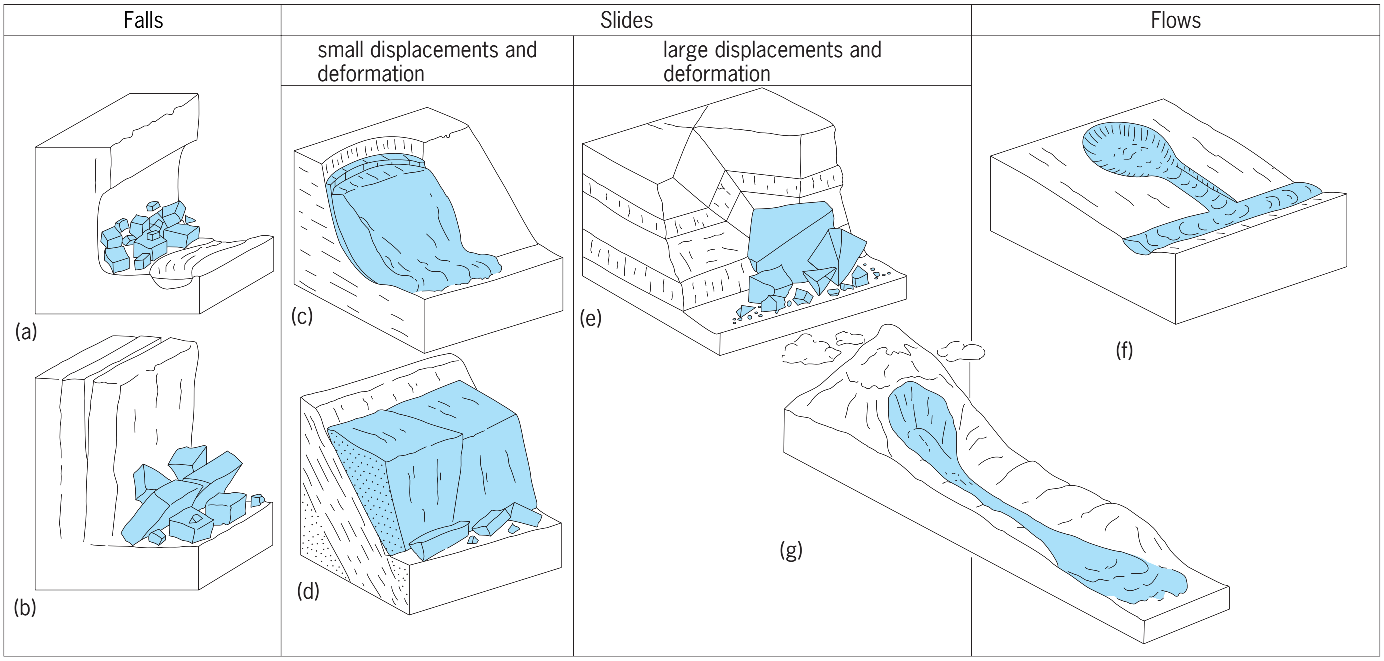


Fig. 1. Types of landslides: (*a*) fall (by undercutting); (*b*) fall (by toppling); (*c*) slump; (*d*) glide; (*e*) rockslide; (*f*) earthflow; and (*g*) rockslide that develops into flow (avalanche). Dimensions may vary considerably from one type to another. The materials involved are rock, soil, and debris.

began as rockslides. A landsliding event may contain more than one type of slide. Thus a rock glide may deform more at its toe and develop a rockslide there. Slumps often exhibit a flowage of material at their toes.

Flows. The internal displacement and the shape of the so-called flows resemble those of viscous fluids,the husaalthough some of them may contain negligible amounts of water. Slip surfaces are not discernible either within the mass or at its base (Fig. 1*f*). Relatively weak and wet masses of shale, weathered rock, and soil may move in the form of debris flows and earthflows; water-soaked soils or weathered rock may displace as mudflows. Some very wet flows may attain large sizes and high velocities and may be able to engulf and displace large rock blocks or human-made structures for miles down their paths.

When extremely large ($0.35–35 \times 10^9$ or $0.01–1 \times 10^9$) masses of rock, ice, and snow tear off the side of a high mountain, the movement is first of the rockslide type (Fig. 1*g*). In some instances, these collossal landslides may be temporarily airborne and hence become rockfalls. In many cases, the gigantic moving masses break down into progressively smaller fragments and engulf snow, ice, and mud to form mixtures that flow very rapidly for long distances and are known as debris avalanches. Some rock avalanches contain very little water. The large runout of relatively dry rock avalanches has been explained as a function of the cushioning effect of a layer of pressurized air trapped at the sole of the mass. The apparent fluidity and the ability of these large rockslides and avalanches to conform to the topographic irregularities of their path, however, may be simply a result of a lack of internal stiffness which results from the breakdown of the original rock mass into progressively smaller blocks until their size is dwarfed by the dimensions of the entire moving mass. Internal collisions allow the transmission of kinetic energy from block to block throughout the mass and create a grain-type of flow in a dispersion, similar to the flow of dry sand. Water vaporization by frictional heat at the base of large rockslides and avalanches can provide a very effective means of reducing base friction and promoting large and violent runouts.

The world's best-known and most devastating landslides have been of the rockslide/avalanche type. In the Frank rockslide of 1903 in southern Alberta, 1×10^9 ft^3 (30×10^6 m^3) of limestone moved along a bedding plane dipping toward the valley, destroyed the southern end of the town of Frank, continued up the opposite side of the valley, and came to rest 20 ft (6 m) above the valley floor. The Gros Ventre rockslide/avalanche of 1925 in northwestern Wyoming was the largest landslide in North America. It involved 1.4×10^9 ft^3 (40×10^6 m^3) of poorly cemented sandstone and was caused by a combination of the progressive leaching of the cement in the sandstone, pore-water pressures caused by heavy rains, and one or more earthquake shocks. The Madison Canyon rockslide (**Fig. 2**) of 1959 in Montana (near Yellowstone Park), and the Hope rockslide of 1965 in the Cascade Mountains of southern British Columbia were also triggered by earthquake shocks. The Huascarán avalanches of 1962 and 1970 in central Peru, were among the most catastrophic natural disasters in historical times. In both events, portions of the side of Mount Huascarán tore off and moved approximately 16 km along a path of destruction. The 1962 event (volume approximately 350×10^6 ft^3 or 10×10^6 m^3; velocity approximately 36 mi/h or 60 km/h) buried the town of Ranrahirca and killed 4000 people. The 1970 event (volume approximately 3.5×10^9 ft^3 or 100×10^6 m^3); velocity approximately 102 mi/h or 170 km/h) was earthquake-triggered, covered the town Yungay under 30 ft (10 m) of mud and rock blocks, and claimed 18,000 lives.

Fig. 2. Madison Canyon rockslide showing slide scar on left wall and slide debris damming the Madison River, forming new lake, in foreground. (*J. R. Stacy*, *USGS*)

Landslides and engineering. Mining and civil engineering works have induced myriads of landslides, a few of them of a catastrophic nature. Open-pit mines and road cuts create very high and steep slopes, often quite close to their stability limit. Local factors (weak joints, fault planes) or temporary ones (surges of water pressure inside the slopes, earthquake shocks) induce the failure of some of these slopes. The filling of reservoirs submerges the lower portion of natural, marginally stable slopes or old landslides. Water lowers slope stability by softening clays and by buoying the lowermost portion, or toe, of the slope. Slope toes often resist and retain the unstable upper portions of slopes; thus lighter (buoyant) toes can be easily pushed toward the reservoir.

Two well-known landslides that occurred in Europe were triggered by human activities. The Elm avalanche of 1881, which destroyed part of the town of Elm in the Swiss Alps and killed 115 people, was triggered by slate-quarrying activity at the foot of a high mountain side. The Vaiont rockslide of 1963 in northern Italy was even more catastrophic. Filling of the Vaiont dam reservoir induced the violent sliding of 8.5×10^9 ft^3 (240×10^6 m^3) of sedimentary rocks into the reservoir; the reservoir water was displaced by the slide material, overtopped the dam (which withstood the event), and rushed down Vaiont gorge to the Piave valley, where it claimed 2600 lives.

Advances in soil and rock engineering have improved the knowledge of slope stability and the mechanics of landsliding. Small and medium-sized slopes in soil and rock can be made more stable. Remedial measures include lowering the slope angle, draining the slope, using retaining structures, compressing the slope with rock bolts or steel tendons, and grouting. Engineering geologists often are able to recognize areas with marginally stable slopes; engineering activity in such areas is then avoided, or proper measures are undertaken. One exception to the ability to recognize a potentially disastrous landslide was the mud flow that killed 128 workers at the Guavio hydroelectric project, east of Bogotá, Colombia, on July 28, 1983. The mud flow moved at great speed, involved colluvium and highly weathered rock (saprolite), and was preceded by record rains. Less catastrophic slides have continued to take place in California and other slide-prone areas around the world, chiefly in response to heavy rains. *See* ENGINEERING GEOLOGY; EROSION; SOIL MECHANICS.

Alberto S. Nieto

Bibliography. M. J. Crozier, *Landslides*: *Causes*, *Consequences and Environment*, 1986; R. L. Schuster and R. J. Krizek (eds.), *Landslides*: *Analysis and Control*, NAS-NAC Spec. Rep. 176, 1978; E. F. S. Sharpe, *Landslide and Related Phenomena*, 1938, reprint 1960; J. E. Slosson, A. G. Keene, and J. A. Johnson (eds.), *Landslides: Landslide Migration*, 1993; B. Voight (ed.), *Rockslides and Avalanches*, pt. 1, 1978, pt. 2, 1980.

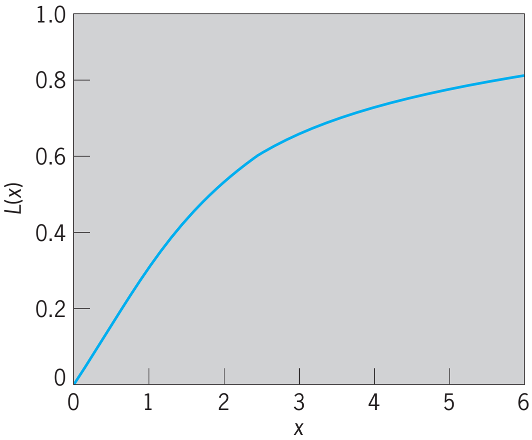
Langevin function

A mathematical function which is important in the theory of paramagnetism and in the theory of the dielectric properties of insulators. The analytical expression for the Langevin function (see **illus.**) is shown in the equation below. If $x < 1$, $L(x) \cong x/3$.

$$L(x) = \coth x - \frac{1}{x}$$

The paramagnetic susceptibility of a classical (non-quantum-mechanical) collection of magnetic dipoles is given by the Langevin function, as is the polarizability of molecules having a permanent electric dipole moment. In the quantum-mechanical treatment of paramagnetism the Langevin function is replaced by the Brillouin function. For further discussion *see* PARAMAGNETISM; PERMITTIVITY.

Elihu Abrahams; F. Keffer



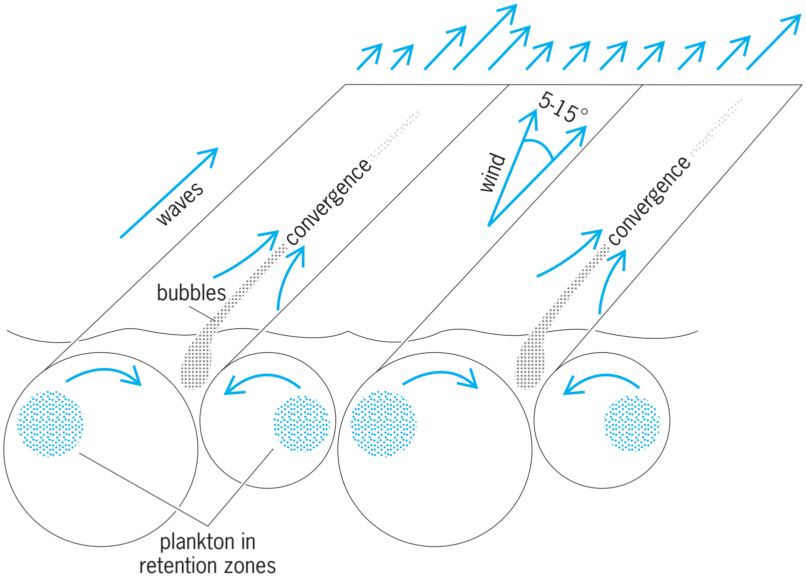
Plot of the Langevin function.

Langmuir circulation

A form of motion found in the near-surface water of lakes and oceans under windy conditions. When the wind is stronger than 5–8 m/s (10–15 knots), streaks of bubbles, seaweed, or flotsam form lines running roughly parallel to the wind, called windrows. Windrows are seen at one time or another on all bodies of water, from ponds to oceans. In the 1920s, Irving Langmuir hypothesized that they are produced by convergences in the water rather than by a direct action of the wind. Langmuir proposed that as the surface water is blown downwind it moves in a spiral fashion, first angling toward the streaks along the surface, next sinking to some depth, then diverging out from under the streaks, and finally rising again in between the streaks (see **illus.**). In a series of observations and experiments conducted in the North Atlantic and on Lake George in New York, he was able to confirm this basic form of the circulation.

In the ocean, the downwelling under windrows can be strong enough to pull bubbles, seaweed, and other buoyant particles down tens to hundreds of meters below the surface. The downward motion is eventually halted by a subtle increase in the water density with depth, associated with colder temperatures and/or higher salinity. The mixed surface layer typically spans most or all of the depth that light penetrates (the euphotic zone). This layer acts like the skin of the sea, through which heat, water vapor, oxygen, carbon dioxide, and all other materials must pass as they enter and leave the air and the sea. The mixing at the bottom of this layer also brings up nutrients and colder water from below.

The existence of an organized pattern of motion invites the notion that marine organisms have evolved to exploit it. For example, A. H. Woodcock noted that most Portuguese man-of-war (*Physalia*) in the North Atlantic are constructed so that they "sail" about 45° to the left of the wind (looking downwind). He speculated that a competitive advantage would be gained if Langmuir circulation were asymmetric (see illus.). Also, H. Stommel argued that particles that sink slowly (such as most phytoplankton) would, upon reaching the lower portion of the rolls, be advected toward the upwelling zones, between streaks, where they would be brought upward by the flow until they diverged away from the upwelling zone. He showed that particles can be trapped in retention zones, which depend on the sinking or rising rate. Sinking particles are trapped toward the upwelling region, while rising ones (such as bubbles) are trapped closer to the downwelling region under the streaks. Thus, the sinking of phytoplankton places them in the upwelling zones where nutrients are more readily available.



Langmuir circulation consists of a set of alternating rolls of water nearly aligned with the wind. Bubbles, seaweed, and flotsam gather into windrows where the surface water converges. Water beneath these windrows flows downward, drawing these materials to depths as great as 100 m or more. These downwelling zones tend to be narrow and intense compared to the broader, gentler upwelling in between. The rolls may be asymmetric, with stronger flows to the right of the wind at the surface (in the Northern Hemisphere). Details of the lower part of the motion are not yet clear.

In the late 1970s a workable mechanism for Langmuir circulation was deduced. A major clue came from laboratory studies by A. J. Faller and E. A. Caponi that demonstrated the need for both surface waves and wind-driven surface shear (downwind flow that is directly forced by the wind, so it is maximal at the surface and decreases rapidly with depth). Rigorous presentations of this mechanism were given by A. D. D. Craik and S. Leibovich. In brief, given a current pattern that resembles a downwind-directed jet (as might be generated by a large breaking wave), a wave-current interaction forces the surface water to flow from the sides toward the center of that jet. Then, since the surface water is accelerating under the wind stress, by the time it gets to the jet, it is moving downwind faster than the surrounding water. Thus, the jetlike feature becomes stronger and a positive feedback cycle is complete. The initially weak streak grows rapidly until some other influence, such as the development of a large enough density jump at the bottom of the mixed layer, arrests the growth.

The general characteristics of Langmuir circulation as known today are:

1. *Spacing*. In general, larger-scale streaks are associated with stronger current variations, up to some limiting size. The largest scale is related to the mixed layer depth (alluded to above), and the spacing between windrows is two to three times this depth. Since a pair of rolls is contained between each pair of windrows, the rolls themselves are about as deep as they are wide.

2. *Orientation*. Although the streaks are roughly aligned with the wind, there is a systematic bias in open ocean conditions, probably due to the effects of the Earth's rotation. In the Northern Hemisphere, oceanic streaks tend to align some 5–15° to the right of the wind (reversed in the Southern Hemisphere).

3. *Length, degree of organization*. The length of the streaks (in the alongwind direction) seems to vary widely, relative to the distance between them. In open ocean storms, the length may be only a few times the spacing. In contrast, for shallow flat-bottom lakes, the streaks can be quite regular, with the streaks extending alongwind distances many hundreds times the crosswind streak spacing. The streak length also relates to how long the features remain stable. In ocean storms, the streaks evolve rapidly compared to the time it would take for water to make a circuit around the roll (from surface to deep to surface again), while the long regular streaks of a shallow lake hardly vary in time.

4. *Strength*. Leaving aside the mean flow associated with wind (about 3% of the windspeed at the surface), the surface water velocities associated with Langmuir circulation are generally about 1% of the windspeed. However, this estimate can be off by a factor of 3 for reasons that remain unknown. Recent detailed examinations of wind and wave data suggests that the surface currents are more closely tied to the waves than the wind, although these are hard to separate since the waves are typically generated by the wind. But the uncertainty by a factor of 3 remains.

Dispersion of materials at the sea surface, such as oil, seaweed, flotsam, and life rafts, depends on the Langmuir circulation. Materials that are swept into the convergences along the streaks will move more rapidly downwind than they otherwise would. In the cross-wind direction, the spreading sidewise from streak to streak depends sensitively on how erratic versus straight and steady the streaks are.

In spite of decades' worth of data and the development of a reasonable driving mechanism, there remains much to learn about Langmuir circulation. In particular, observations to date have concentrated on the surface circulation, so the bottom portion of the motion has not been well defined.

Jerome A. Smith

Bibliography. A. D. D. Craik, The generation of Langmuir circulation by an instability mechanism, *J. Fluid Mech.*, 81:209–223, 1977; A. J. Faller and E. A. Caponi, Laboratory studies of wind-driven Langmuir circulations, *J. Geophys. Res.*, 83:3617–3633, 1978; I. Langmuir, Surface motion of water induced by wind, *Science*, 87:119–123, 1938; S. Leibovich, Convective instability of stably stratified water in the ocean, *J. Fluid Mech.*, 82:561–581, 1977; J. A. Smith, Evolution of Langmuir circulation during a storm, *J. Geophys. Res.*, 103:12649–12668, 1998; H. Stommel, Trajectories of small bodies sinking slowly through convection cells, *J. Mar. Res.*, 8:24–29, 1949; S. A. Thorpe, On the meandering and dispersion of a plume of floating particles caused by Langmuir circulation and a mean current, *J. Phys. Oceanog.*, 25:685–690, 1995; A. H. Woodcock, A theory of surface water motion deduced from the wind-induced motion of *Physalia*, *J. Mar. Res.*, 5:196–205, 1994.

Language theory

An attempt to formulate the grammar of a language in mathematical terms. Language theory is an important area of linguistics and computer science. Formal language theory was initiated in the mid-1950s in an attempt to develop theories on natural language acquisition. This theory, and in particular context-free grammars, was found to be relevant to the languages used in computers. Interest in the relationship between abstract languages and automata theory began with a seminal paper by S. C. Kleene in 1956, in which he characterized the languages in which membership of a sentence could be decided by a finite-state machine. Formal or abstract languages are based on the mathematical notion of a language as defined by Noam Chomsky around 1956. To understand this concept, we may begin by defining what a language is. The *New Oxford American Dictionary* defines it as "the method of human communication, either spoken or written, consisting of the use of words in a structured and conventional way." However, this definition is too vague to use as a building block of a language theory. To formalize the notion of an abstract language, it is necessary to introduce some preliminary definitions. An alphabet, vocabulary, or the set of terminals, denoted by Σ, is any finite, nonempty set of indivisible symbols. For example, the binary alphabet has only two symbols. This set is generally represented as $\Sigma = \{0, 1\}$. A word or string, over a particular alphabet, is a finite sequence of symbols of the alphabet. In mathematical terms, a typical word, x, can be written as $x = a_1, a_2, a_3, \ldots, a_k$ where $k \geq 0$, $a_i \in \Sigma$ for $1 \leq i \leq k$. Notice that if $k = 0$, the word is called the null word or empty word and is denoted by Λ. For example, using the binary alphabet we can form the words $x = 0010$ and $y = 010$. Given a word, x, the number of occurrences of symbols of a given alphabet in the word is called the length of the word and is denoted by $|x|$. According to this definition, the length of the words $|x| = |0010|$ and $|y| = |010|$ are 4 and 3, respectively. For a particular alphabet, a sentence is a finite sequence of words. *See* AUTOMATA THEORY; LINGUISTICS.

For a given alphabet, an abstract language is the set of all valid sentences of the language. A sentence may be considered valid if it satisfies a finite set of formation or production rules that define the grammar of the language. This definition of language can be applied to any natural or computer language. In a binary language, for instance, if we consider all the sequences of binary symbols of length 4 or less as valid words, then the words $x = 0010$ and $y = 010$ belong to this language. Sentences for a given language can be generated through a grammar. A grammar G can be formally defined as $G = (N, \Sigma, P, S)$, where N is a finite set of nonterminal symbols, also called variables or syntactical categories. Σ is called the alphabet or set of terminal symbols. Σ and N do not have any elements in common. That is, $\Sigma \cap N = \emptyset$. P is a finite set of production or replacement rules. Replacement rules have the form $\alpha \rightarrow \beta$, where α, β are strings. This rule indicates that an instance or occurrence of α can be replaced by an instance or occurrence of β. S is a distinguished element of N called the starting symbol. For example, let us consider a formal minigrammar G that generates four English sentences. This grammar has the following productions or replacement rules:

1. <sentence> → <subject> <verb> <object>
2. <subject> → <article> <noun>
3. <object> → <article> <noun>
4. <verb> → SEES
5. <article> → THE
6. <noun> → BOY
7. <noun> → GIRL

From this set of production rules, we can distinguish the elements that make up the grammar. The set of nonterminal symbols, N, is formed by the entities enclosed in pointed brackets. That is, N = {<sentence>, <subject>, <object>, <verb>, <article>, <noun>}. The alphabet Σ has four elements, namely, Σ = {SEES, THE, BOY, GIRL}. Each of the elements of the alphabet should be considered an indivisible unit. The replacement rules indicate how sequences of words in the language can be put together to generate sentences. For the given set of replacement rules, the interpretation of the first production is that a sentence is formed by a <subject> followed by a <verb> followed by an <object>. Similar interpretations can be given to rules 1 through 3. Rules 4 and 5 indicate that SEES, THE, BOY, and GIRL are instances of a <verb>, <article>, or <noun>, respectively. When there is more than one choice for a particular entity, such as <noun>, the entity can be replaced by the right-hand sides of one of the production rules. In this case, <noun> can be replaced by either BOY or GIRL. The starting symbol of this grammar is, by convention, the element of N that appears at the left of the symbol → of the first production. In this case, <sentence> is the starting nonterminal. One sentence generated by this grammar is THE GIRL SEES THE BOY. The sequence of replacement rules used to generate or derive this sentence is as follows:

<sentence> → <subject> <verb> <object>.
<sentence> → <article> <noun> <verb> <object>.
(using production 2 to replace <subject>)
<sentence> → THE <noun> <verb> <object>.
(using production 5 to replace <article>)
<sentence> → THE GIRL <verb> <object>.
(using production 7 to replace <noun>)
<sentence> → THE GIRL SEES <object>.
(using production 4 to replace <verb>)
<sentence> → THE GIRL SEES <article> <noun>
(using production 2 to replace <subject>)
<sentence> → THE GIRL SEES THE <noun>
(using production 5 to replace <article>)
<sentence> → THE GIRL SEES THE BOY.
(using production 6 to replace <noun>)

Notice that all the elements of the sentence THE

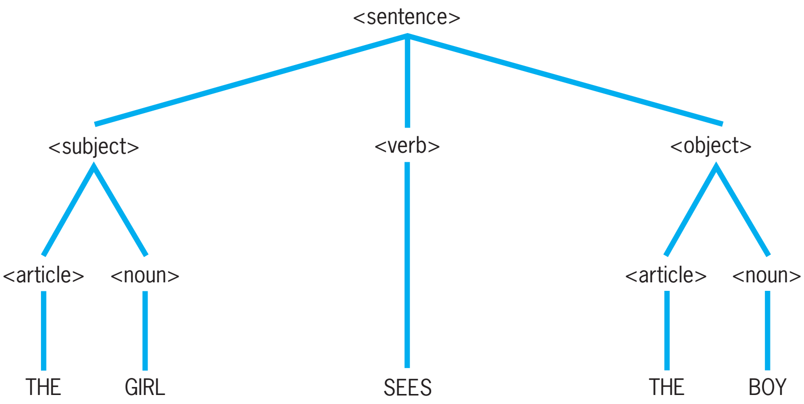


Fig. 1. Derivation tree of the sentence THE GIRL SEES THE BOY.

GIRL SEES THE BOY belong to the set of terminals. At each step of the derivation, a nonterminal on the right-hand side of a production was replaced by another sequence of nonterminal symbols or by a single terminal. The three additional sentences that can be generated using the production rules of this grammar are: THE BOY SEES THE GIRL, THE BOY SEES THE BOY, THE GIRL SEES THE GIRL. These four sentences form the language generated by the grammar. As this example shows, it is not necessary to explicitly indicate the elements of N, Σ, P, and S of the definition of a grammar since all these elements can be derived directly from the set of production rules. **Figure 1** shows a derivation tree of the sentence generated before.

Types of grammars. Grammars can be classified according to the nature of their productions. If G = (N, Σ, P, S) and for every production $\alpha \rightarrow \beta$ in P, $|\beta| \geq |\alpha|$, the grammar is said to be context-sensitive or type 1. Some authors require that the productions of this type of grammar be of the form $\alpha_1 A \alpha_2 \rightarrow \alpha_1 \beta \alpha_2$, where α_1, α_2, and β are strings, $\beta \neq \Lambda$, and $A \in N$. This restriction, which does not alter the language generated by the grammar, motivates the term context-sensitive because A can be replaced by β only within the context of α_1 and α_2 according to $\alpha_1 A \alpha_2 \rightarrow \alpha_1 \beta \alpha_2$.

If the productions $\alpha \rightarrow \beta$ of the grammar G are such that $|\alpha| = 1$ and α is a single element of N and β is a string of terminals and/or nonterminals, then the grammar is a context-free grammar or type 2. Context-free grammars and the language that they generate are very important in computer science since most of the computer languages are generated by this type of grammar.

Acceptors, generators, and translators. The relationship between abstract languages and machines can be established through three types of automata: acceptors, generators, and translators. An acceptor is a deterministic finite-state machine which, starting in a predefined initial state, accepts or rejects sequences of input symbols from an alphabet. We say that a finite-state machine operates deterministically if its responses depend solely on its initial state and the input sequence presented. In simple terms, we can picture an acceptor as a machine connected to a lamp which can be turned on and off. As the machine processes each of the symbols of the input sequence, the machine turns the lamp on or off. If the machine turns the lamp on after processing the last symbol of the input sequence, the machine accepts the input sequence; otherwise, the machine rejects the sequence.

A generator is a nondeterministic finite-state machine. When the generator is started from its initial state, it produces a sequence of symbols from a given alphabet known as the output alphabet. It is assumed that the machine will act according to chance producing different sequences each time it is operated. The language of the machine is the set of all sequences that it will ever produce.

A translator is a machine that after processing an input sentence (made up of symbols of an input alphabet) translates the sentence into another sentence (made of symbols of an output alphabet). Translators are important not only in the translation of natural languages (for example, from English

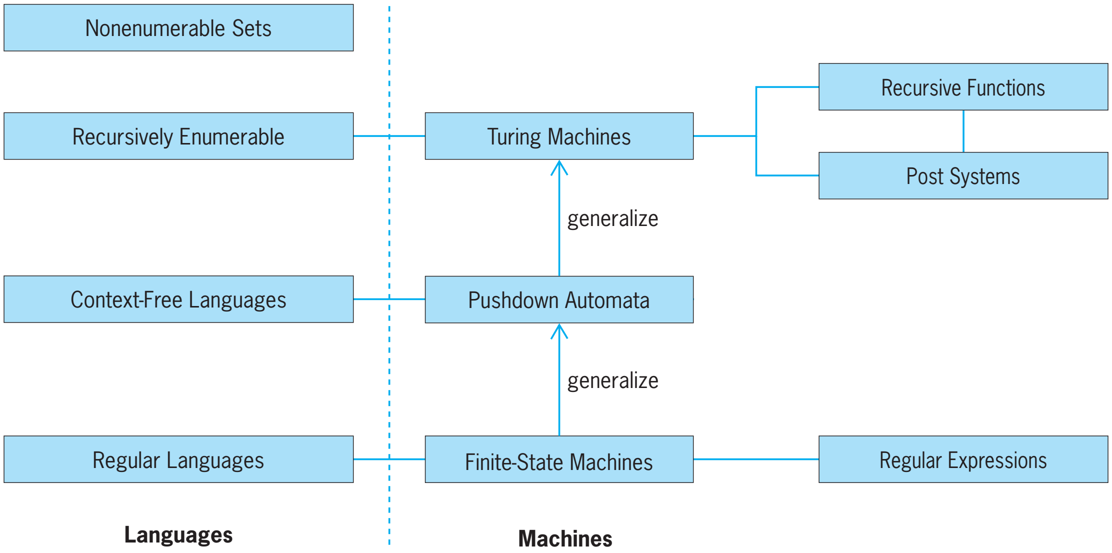


Fig. 2. Hierarchy of abstract machines and languages. (*After P. Denning et al., Machines, Languages, and Computation, 1978*)

to Spanish) but also in the translation of computer languages (for example, from C++ to assembly or machine language). *See* NATURAL LANGUAGE PROCESSING; PROGRAMMING LANGUAGES.

Figure 2 shows a hierarchy of automata and the corresponding hierarchy of languages that they define. Although some of the topics mentioned in this hierarchy are out of the scope of this article, they are included here for completeness.

Ramon A. Mata-Toledo

Bibliography. A. V. Aho et al., *The Theory of Parsing, Translation, and Compiling*, vol. 1, 1972 J. E. Hopcroft et al., *Introduction to Automata Theory, Languages, and Computation*, 2d ed., 2000; J. Martin, *Introduction to Languages and the Theory of Computation*, 2002.

Lanolin

A soft, waxy material derived from the greasy coating on raw wool. In the processing of wool, the fleece is scoured with an aqueous alkaline solution to remove debris, wax, water-soluble material, and free acids. The insoluble fraction, which constitutes about 15–20% of the weight of the original wool, is crude lanolin.

Lanolin is a very complex mixture of esters, similar to the skin lipids of birds and other animals. Analysis is carried out after saponification, which gives equal weights of acids and alcohols. The acids consist of 50–60% saturated fatty acids with chains up to C_{38}–C_{40}, comprising mostly isostructures or anteisostructures with a methyl group that is one or two carbons removed from the end of the chain [$CH_3CH(CH_3)$— or $CH_3CH_2CH(CH_3)$—]. Another group (30–35%) comprises α-hydroxy acids and a small amount of ω-hydroxy acids. *See* ESTER.

The alcohols from lanolin are of two types. About 15% are monohydroxy compounds with the same chain structures as the acids, and a minor amount of 1,2-diols. The major alcohol components are 35% cholesterol and 35–40% lanosterol. The latter is a key intermediate in the biosynthesis of steroids from squalene. *See* CHOLESTEROL; SQUALENE; STEROID.

The supply of lanolin depends directly on the level of wool production; the consumption of lanolin and derived products in the United States is about 1×10^7 lb (4.5×10^6 kg) per year. Crude lanolin from the wool-scouring process can be refined to a colorless, odorless material (USP grades). A particularly useful property of lanolin is the formation of stable water-in-oil emulsions containing up to 25% water (hydrous lanolin). Major uses are as an emollient and skin moisturizer in lotions and cosmetic products, and in medicinal ointments. Unrefined lanolin has some use in inks and as a corrosion and rust preventative. *See* FAT AND OIL; WOOL.

James A. Moore

Lanthanide contraction

The name given to an unusual phenomenon encountered in the rare-earth series of elements. The radii of the atoms of the members of this series decrease slightly as the atomic number increases. Starting with element 58 in the periodic table, the balancing electron fills in an inner incomplete 4*f* shell as the charge on the nucleus increases. According to the theory of atomic structure, this shell can hold 14 electrons; so starting with element 58, cerium, there are 14 true rare earths. Lanthanum has no electrons in the 4*f* shell, cerium has 1, and lutetium, 14. The 4*f* electrons play almost no role in chemical valence; therefore, all rare earths can have three electrons in their valence shell and they all exist as trivalent ions in solution. As the charge on the nucleus increases across the rare-earth series, all electrons are pulled in closer to the nucleus so that the radii of the rare-earth ions decrease slightly as the compounds go across the rare-earth series. Any given compound of the rare earths is very likely to crystallize with the same structure as any other rare earth. However, the lattice parameters become smaller and the crystal denser as the

Atomic and ionic radii of rare-earth metals

Element	Radius of M^{3+}[*], nm	Metal crystal structure[†]	Metallic radii, nm	
Sc		hcp	0.16545[‡]	0.16280[§]
Y		hcp	0.18237	0.17780
La	0.1061	hcp	0.18852	0.18694
Ce	0.1034	fcc	0.18248	
Pr	0.1013	hcp	0.18363	0.18201
Nd	0.0995	hcp	0.18290	0.18139
Pm	0.0979			
Sm	0.0964	rhom-hcp	0.18105	0.17943
Eu	0.0950	bcc	0.1994	
Gd	0.0938	hcp	0.18180	0.17865
Tb	0.0923	hcp	0.18005	0.17626
Dy	0.0908	hcp	0.17952	0.17515
Ho	0.0894	hcp	0.17887	0.17428
Er	0.0881	hcp	0.17794	0.17340
Tm	0.0869	hcp	0.17688	0.17237
Yb	0.0858	fcc	0.19397	
Lu	0.0848	hcp	0.17516	0.17171

[*] Trivalent rare-earth ion.

[†] hcp, hexagonal close-packed; fcc, face-centered cubic; rhom, rhombic; bcc, body-centered cubic.

[‡] All values in this column calculated from atoms in basal plane.

[§] All values in this column are for radii between layers.

compound proceed across the series. This contraction of the lattice parameters is known as the lanthanide contraction. For many compounds the lattice parameters decrease only partway across the series, and when the contraction has progressed to that point, a new crystalline form develops. Frequently, both crystalline forms can be observed for a number of the elements. For this reason, the rare-earth series is of particular interest to scientists because many of the parameters determining the properties of a substance can be kept constant while the lattice spacings can be varied in small increments across the series.

The atomic and ionic radii of atoms are not clearly defined. The atoms can be polarized by the neighboring atoms and there is no clear-cut boundary between the electrons associated with one atom and another. Therefore, the atomic radii will vary somewhat from compound to compound, and the absolute values depend on the method of calculation. However, if most of the parameters are assumed constant, and the difference in lattice parameters in the rare-earth crystalline series is attributed to the rare-earth ion or atom, the lanthanide contraction becomes evident. Although scandium and yttrium are not members of this series, the information is usually wanted at the same time and is given for completeness. The atomic radii of the trivalent ion and the metal atoms are given in the **table**. *See* PERIODIC TABLE; RARE-EARTH ELEMENTS.

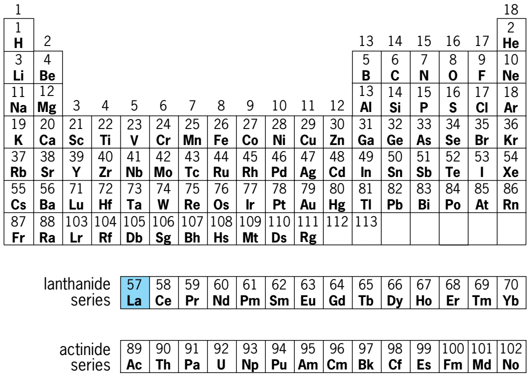
Frank H. Spedding

Lanthanum

A chemical element, La, atomic number 57, atomic weight 138.91. Lanthanum, the second most abundant element in the rare-earth group, is a metal. The naturally occurring element is made up of the isotopes ^{138}La, 0.089%, and ^{139}La, 99.91%. ^{138}La is a radioactive positron emitter with a half-life of 1.1×10^{11} years. The element was discovered in 1839 by C. G. Mosander and occurs associated with other rare earths in monazite, bastnasite, and other minerals. It is one of the radioactive products of the fission of uranium, thorium, or plutonium. Lanthanum is the most basic of the rare earths and can be separated rapidly from other members of the rare-earth series by fractional crystallization. Considerable quantities of it are separated commercially, since it is an important ingredient in glass manufacture. Lanthanum imparts a high refractive index to the glass and is used in the manufacture of expensive lenses. The metal is readily attacked in air and is rapidly converted to a white powder. Lanthanum becomes a superconductor below about 6 K (−449°F) in both the hexagonal and face-centered crystal forms. *See* RARE-EARTH ELEMENTS.

Frank J. Spedding

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; K. A. Gschneidner Jr., J.-C. Bünzli, and V. K. Pecharsky (eds.), *Handbook on the Physics and Chemistry of Rare Earths*, 2005.



1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1 **H**																	2 **He**
3 **Li**	4 **Be**											5 **B**	6 **C**	7 **N**	8 **O**	9 **F**	10 **Ne**
11 **Na**	12 **Mg**											13 **Al**	14 **Si**	15 **P**	16 **S**	17 **Cl**	18 **Ar**
19 **K**	20 **Ca**	21 **Sc**	22 **Ti**	23 **V**	24 **Cr**	25 **Mn**	26 **Fe**	27 **Co**	28 **Ni**	29 **Cu**	30 **Zn**	31 **Ga**	32 **Ge**	33 **As**	34 **Se**	35 **Br**	36 **Kr**
37 **Rb**	38 **Sr**	39 **Y**	40 **Zr**	41 **Nb**	42 **Mo**	43 **Tc**	44 **Ru**	45 **Rh**	46 **Pd**	47 **Ag**	48 **Cd**	49 **In**	50 **Sn**	51 **Sb**	52 **Te**	53 **I**	54 **Xe**
55 **Cs**	56 **Ba**	71 **Lu**	72 **Hf**	73 **Ta**	74 **W**	75 **Re**	76 **Os**	77 **Ir**	78 **Pt**	79 **Au**	80 **Hg**	81 **Tl**	82 **Pb**	83 **Bi**	84 **Po**	85 **At**	86 **Rn**
87 **Fr**	88 **Ra**	103 **Lr**	104 **Rf**	105 **Db**	106 **Sg**	107 **Bh**	108 **Hs**	109 **Mt**	110 **Ds**	111 **Rg**	112	113					

lanthanide series	57 **La**	58 **Ce**	59 **Pr**	60 **Nd**	61 **Pm**	62 **Sm**	63 **Eu**	64 **Gd**	65 **Tb**	66 **Dy**	67 **Ho**	68 **Er**	69 **Tm**	70 **Yb**
actinide series	89 **Ac**	90 **Th**	91 **Pa**	92 **U**	93 **Np**	94 **Pu**	95 **Am**	96 **Cm**	97 **Bk**	98 **Cf**	99 **Es**	100 **Fm**	101 **Md**	102 **No**

Laplace transform

An integral extensively used by P. S. Laplace in the theory of probability. In simplest form it is expressed as Eq. (1). It is thought of as transforming the determining function $\phi(t)$ into the generating function $f(s)$. The variable t is real, the variable s may be real or complex, $s = \sigma + i\tau$. As an example, if $\phi(t) = 1$ the integral converges for $\sigma > 0$, and $f(s) = 1/s$.

$$f(s) = \int_0^\infty e^{-st}\phi(t)\,dt \qquad (1)$$

The Laplace transform is used for the solution of differential and difference equations, for the evaluation of definite integrals, and in many branches of abstract mathematics (functional analysis, operational calculus, and analytic number theory).

Method. Extensive tables of Laplace transforms exist, and these are used as any table of integrals. To show how a differential equation may be solved, two excerpts (A and B) from such a table can be used.

A. $f(s) = 1/(s - a)$	$\phi(t) = e^{at}$
B. $f(s) = 1/(s^2 + 1)$	$\phi(t) = \sin t$

Suppose it is required to find a solution $y(t)$ of Eq. (2) such that $y(0) = 1$, $y'(0) = 2$. Denote the Laplace transform of the unknown function $y(t)$ by $Y(s)$. Integration by parts gives Eq. (3) on the assumption that the integrated part is zero at $t = \infty$. Applying the Laplace transform to Eq. (2) and using A for the right-hand side, one obtains Eq. (4). The differential equation has become an algebraic one, whose

$$y''(t) + y(t) = 2e^t y'' = \frac{d^2y}{dt^2}, y' = \frac{dy}{dt} \qquad (2)$$

$$\int_0^\infty e^{-st}y''(t)\,dt = -y'(0) - y(0)s + s^2\int_0^\infty e^{-st}y(t)\,dt = -2 - s + s^2Y(s) \qquad (3)$$

$$-2 - s + s^2Y(s) + Y(s) = \frac{2}{s-1} \qquad (4)$$

solution is Eq. (5). However, a further use of the table

$$Y(s) = \frac{1}{s-1} + \frac{1}{s^2+1} \qquad (5)$$

shows that the Laplace transform of $y(t) = e^t + \sin t$ is precisely the right-hand side of Eq. (5). Assuming uniqueness, one has thus obtained the required solution. Because its properties can be checked directly, the unproved assumptions need not be verified.

This example illustrates the general method. The unknown function is taken as the determining function and the Laplace transform is applied to the differential (or difference) equation. There results an equation with the generating function as unknown, and this must be solved. Finally the determining function must be determined from the generating function, either from tables or by use of an inversion formula. In general, if the original differential equation is partial in any number of independent variables, one application of the Laplace transform reduces the number of these variables by one. If the equation was ordinary (one independent variable), the transformed equation is algebraic, as in the above example.

Properties. Here are the fundamental properties of the Laplace transform:

I. There exists a number σ_c (perhaps $+\infty$ or $-\infty$) called the abscissa of convergence such that the integral in Eq. (1) converges for $\sigma > \sigma_c$, diverges for $\sigma < \sigma_c$. That is, the region of convergence is a half-plane (a half-line if s is real).

II. The generating function is holomorphic for $\sigma > \sigma_c$.

III. The determining function is uniquely determined by the generating function. (Ambiguity is possible only on sets of measure zero.)

IV. The product of two generating functions is in general a generating function. Thus, if Eq. (1) holds for two pairs of functions $f_1(s)$, $\phi_1(t)$ and $f_2(s)$, $\phi_2(t)$, then the product $f_1(s)f_2(s)$ is the transform of the convolution

$$\phi_1(t)^*\phi_2(t) = \int_0^t \phi_1(u)\phi_2(t-u)\,du$$

As was evident in the above example, it is very important to be able to derive the determining function $\phi(t)$ from the generating function $f(s)$. This is especially true when tables are unavailable or inadequate. Any expression of $\phi(t)$ in terms of $f(s)$ is called an inversion formula. Many are known. The classical one is Eq. (6). Here the integration is along any line

$$\phi(t) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} f(s)e^{st}ds \qquad 0 < t < \infty \qquad (6)$$

$\sigma = c$ of the complex s-plane on which the integral in Eq. (1) converges absolutely. Another inversion which employs the real variable only is Eq. (7). Here

$$\phi(t) = \lim_{k\to\infty} \frac{(-1)^k}{k!} f^{(k)}\left(\frac{k}{t}\right)\left(\frac{k}{t}\right)^{k+1} \qquad (7)$$

$$0 < t < \infty$$

$f^{(k)}(x)$ means the kth derivative of $f(x)$. Equation (7) can be illustrated by the example A above. For that pair $f^{(k)}(x)$ is easily computed and Eq. (7) becomes Eq. (8), a familiar result of calculus.

$$e^{at} = \lim_{k\to\infty}\left(1 - \frac{at}{k}\right)^{-k-1} \qquad (8)$$

See INTEGRAL TRANSFORM.

Generalizations. Certain generalizations of Eq. (1) are in frequent use. The transform shown as Eq. (9)

$$f(s) = \int_{-\infty}^{\infty} e^{-st}\phi(t)\,dt \qquad (9)$$

is called the bilateral Laplace transform. An inversion is still provided by Eq. (6), which now holds for $-\infty < t < \infty$. If one sets $s = iy$ in Eq. (9), the result is Eq. (10). This equation defines $g(y)$ as the Fourier

$$g(y) = f(iy) = \int_{-\infty}^{\infty} e^{-iyt}\phi(t)\,dt \qquad (10)$$

transform of $\phi(t)$. That is, the Laplace transform (9), if considered along a single line, becomes a Fourier transform. By setting $c = 0$ and $s = iy$ in formula (6) one obtains Eq. (11), the classical inversion of the Fourier transform.

$$\begin{aligned}\phi(t) &= \frac{1}{2\pi i}\int_{-i\infty}^{i\infty} f(s)e^{st}\,ds \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} g(y)e^{iyt}\,dy \qquad (11)\end{aligned}$$

Another generalization of Eq. (1) is the Laplace-Stieltjes integral, Eq. (12), where now the integral

$$f(s) = \int_0^{\infty} e^{-st}\,d\alpha(t) \qquad (12)$$

is a Stieltjes integral with respect to the "integrator" function $\alpha(t)$. If $\alpha(t)$ has a derivative $\phi(t)$ the integral (12) becomes the integral (1). On the other hand, if $\alpha(t)$ is a step-function, Eq. (12) reduces to a Dirichlet series, Eq. (13), a type of series of great importance

$$f(s) = \sum_{k=1}^{\infty} a_k e^{-\lambda_k s} \qquad (13)$$

in analytic number theory.

It must not be supposed that one may choose $\phi(t)$ or $f(s)$ arbitrarily in Eq. (1) and expect its mate to exist. For example if $\phi(t) = e^{t2}$ the integral (1) diverges for all s and $\sigma_c = +\infty$. Again if $f(s) = s$, no corresponding determining function $\phi(t)$ exists, since it is easily seen that every generating function must approach a limit as $s \to +\infty$ along the real axis. Hence it is clearly important to know what functions $\phi(t)$ and $f(s)$ may be used in Eq. (1). So far as $\phi(t)$ is concerned the problem is completely solved by the formula, with $\sigma_c > 0$, shown as Eq. (14). The other

$$\sigma_c = \overline{\lim_{t\to\infty}} \frac{\log|\alpha(t)|}{t} \qquad (14)$$

problem has been partially solved by representation theorems for integral transforms, one striking example of which is presented below.

A function $f(s)$ of the real variable s is said to be completely monotonic on $a < s < \infty$ if and only if relations (15) hold.

$$f(s) \geq 0, f'(s) \leq 0, f''(s) \geq 0, f'''(s) \leq 0, \ldots \qquad a < s < \infty \tag{15}$$

Examples are $f(s) = 1$, $f(s) = 1/(s - a)$, and $f(s) = e^{-s}$. A theorem of S. Bernstein states that $f(s)$ has a representation (12) converging for $s > a$ and with integrator function $\alpha(t)$ nondecreasing if and only if $f(s)$ is completely monotonic for $a < s < \infty$. For example, if $f(s) = 1$, then $\alpha(t) = 1$, $t > 0$, $\alpha(0) = 0$; if $f(s) = 1/(s - a)$, then $\phi(t) = e^{at}$ and $\alpha(t) = (e^{at} - 1)/a$; if $f(s) = e^{-s}$, $\alpha(t) = 0$ for $0 < t < 1$ and $\alpha(t) = 1$ for $1 < t < \infty$. In each case $\alpha(t)$ is nondecreasing, as predicted by Bernstein's theorem. This result is particularly remarkable because the mere signs of the successive derivatives of a function on the real axis determine not only its holomorphic character (property II above) in a half-plane but also its representation in the form of Eq. (12). David V. Widder

Bibliography. R. E. Bellman and R. S. Roth, *The Laplace Transform*, 1984; A. C. Grove, *Introduction to the Laplace Transform and the Z Transform*, 1991; W. R. LePage, *Complex Variables and the Laplace Transform for Engineers*, 1961, reprint 1980.

Laplace's differential equation

Laplace's equation in two independent variables x and y is given as Eq. (1) and is of central importance in both pure mathematics and mathematical physics.

$$\frac{\partial^2 u(x,y)}{\partial x^2} + \frac{\partial^2 u(x,y)}{\partial y^2} = 0 \tag{1}$$

A function $u(x,y)$ having continuous first and second partial derivatives and satisfying Laplace's equation in a neighborhood of a point is called harmonic at that point. If a plane piece of tinfoil has its edges kept at a temperature which varies from point to point but does not change with time, and if the flow of heat in the tinfoil is steady (that is, independent of the time), the temperature $u(x,y)$ at interior points of the foil is harmonic. Likewise Laplace's equation dominates the flow of electricity (the potential is similarly harmonic) and the flow of any incompressible fluid.

Two-dimensional relations. If $f(z) \equiv u(x, y) + iv(x, y)$ is an analytic function, $u(x, y)$ and $v(x, y)$ are conjugate functions and are harmonic; conversely, if $u(x, y)$ is harmonic in a simply connected region D, one may write Eq. (2), where (x_0, y_0) is fixed in D and (x, y) arbitrary in D.

$$v(x,y) \equiv \int_{(x_0,y_0)}^{(x,y)} \left(-\frac{\partial u}{\partial y}\, dx + \frac{\partial u}{\partial x}\, dy \right) \tag{2}$$

It follows from Green's theorem that the integral over a path in D is independent of the path, so $v(x, y)$ is uniquely defined throughout D; the functions $u(x, y)$ and $u(x, y)$ are conjugate in D, and $f(z) \equiv u + iv$ is analytic there. Under these conditions, let C now be a regular Jordan curve in D; if n denotes the interior normal of C, the equation $\partial u/\partial n = -\partial v/\partial s$ follows from the Cauchy-Riemann equations, whence obtains Eq. (3).

$$\int_C \frac{\partial u}{\partial n}\, ds = -\int_C \frac{\partial v}{\partial s}\, ds = -v(x,y)\Big|_C = 0 \tag{3}$$

The first and last members of this equation form the flux theorem, namely that the total flux (of heat if u is temperature) over C is zero. *See* COMPLEX NUMBERS AND COMPLEX VARIABLES.

If $u(x, y)$ is harmonic in the closed disk bounded by the circumference γ, and $f(z)$ the corresponding analytic function, one can take the real parts of both members of the equations expressing Cauchy's integral formula, as in Eqs. (4).

$$z - z_0 = \rho(\cos\theta + i\sin\theta) \tag{4a}$$

$$dz = i(z - z_0)\, d\theta \tag{4b}$$

$$f(z_0) = u(x_0, y_0) + iv(x_0, y_0) = \frac{1}{2\pi i}\int_\gamma \frac{f(z)\, dz}{z - z_0}$$

$$= \frac{1}{2\pi}\int_\gamma f(z)\, d\theta \tag{4c}$$

$$u(x_0, y_0) = \frac{1}{2\pi}\int_\gamma u(x, y)\, d\theta \tag{4d}$$

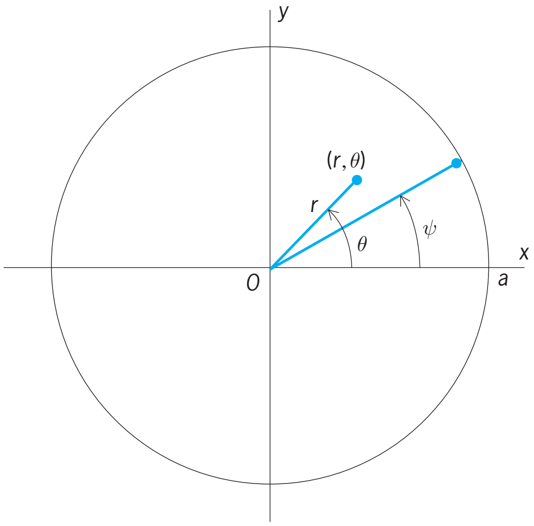
Equation 4*d* expresses Gauss's mean value theorem, that the average of $u(x, y)$ over γ is the value at the center of γ. From this theorem it follows that a function harmonic at a point (x_0, y_0) cannot have a strong local maximum (or minimum) there, and can have a weak local maximum (or minimum) only if identically constant throughout a neighborhood of (x_0, y_0). If $u(x, y)$ is harmonic in a bounded region D, continuous in the corresponding closed region $\overline{D}$, the maximum and minimum of $u(x, y)$ occur on the boundary of D; if a maximum or minimum occurs interior to D, then $u(x, y)$ is identically constant throughout D.

If D is a bounded region with boundary B, and if continuous values $U(x, y)$ are assigned on B, the Dirichlet problem is the problem of determining a function $u(x, y)$ harmonic in D, continuous on $D + B$, equal to $U(x, y)$ on B. If D is a circular region, a Jordan region, or any nonpathological region, the Dirichlet problem has a solution, necessarily (by the absence of nontrivial maxima and minima interior to D) unique. If D is a circular disk of radius a (see **illus.**), the Dirichlet problem for D is solved by Poisson's integral, Eq. (5), using polar coordinates $(r\theta)$

$$u(r, \theta) = \frac{1}{2\pi}\int_0^{2\pi} \frac{(a^2 - r^2)U(\psi)\, d\psi}{a^2 - 2ar\cos(\theta - \psi) + r^2} \tag{5}$$

with pole the center of D. If D is a less elementary region but with smooth boundary B, the Dirichlet problem is solved by Green's formula, Eq. (6), where n indicates the interior normal in this formula.

$$u(x,y) = \frac{1}{2\pi}\int_B U(\xi,\eta)\frac{\partial g}{\partial n}\, ds\, (\xi,\eta) \tag{6}$$



Circular disk *D* of radius *a*.

Green's function $g(x, y; \xi, \eta)$ is harmonic in D except at (x, y), continuous and equal to zero on B, and in the neighborhood of (x, y) has the form $^1/_2$ log $[(\xi - x)^2 + (\eta - y)^2] + g_1(\xi, \eta)$, where $g_1(\xi, \eta)$ is harmonic at (x, y). If the boundary B is not smooth, this formula can be expressed in terms of harmonic measure instead of $(\partial g/\partial n)\, ds$.

Numerous series expansions (for example, Fourier's series) can be used for the solution of the Dirichlet problem for various regions.

n-Dimensional relations. The foregoing remarks apply to Laplace's equation with two independent variables; the facts (but not the methods of proof using analytic functions) apply also in three or more dimensions. Thus, in three dimensions, a point distribution of matter of masses m_k at points (x_k, y_k, z_k) has a potential defined by Eq. (7), which is harmonic except in the points (x_k, y_k, z_k). Except at such points, the force (newtonian law of gravitation) exerted by the distribution on a unit exploratory particle at (x, y, z) has the components $(\partial u/\partial x, \partial u/\partial y, \partial u/\partial z)$, and the component of the force in any direction is the directional derivative of $u(x, y)$ in that direction. *See* POTENTIALS; SPHERICAL HARMONICS.

$$u(x, y, z) \equiv \sum m_k[(x - x_k)^2 + (y - y_k)^2 + (z - z_k)^2]^{-1/2} \qquad (7)$$

Joseph L. Walsh

Bibliography. J. Bliedtner, *Potential Theory*, 1986; J. Kral (ed.), *Potential Theory*, 1988.

Laplacian

The differential operator $\partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$, in which the symbols x, y, z denote the variables of a rectangular cartesian coordinate system. The laplacian is frequently denoted by the symbol ∇^2 (read del square) in accordance with the fact that the laplacian of a scalar function $S(x, y, z)$ is the divergence of the gradient of S, that is, the equation below applies.

$$\frac{\partial^2 S}{\partial x^2} + \frac{\partial^2 S}{\partial y^2} + \frac{\partial^2 S}{\partial z^2} = \nabla \cdot (\nabla S)$$

The laplacian operator is involved in some of the most fundamental equations of mathematical physics, namely, Laplace's equation ($\nabla^2 u = 0$), Poisson's equation, various wave equations (as those of electricity and magnetism, sound, vibrations, the Schrödinger equation of quantum mechanics), and the heat flow and diffusivity equations. *See* CALCULUS OF VECTORS; GAUSS' THEOREM; GRADIENT OF A SCALAR; GREEN'S THEOREM.

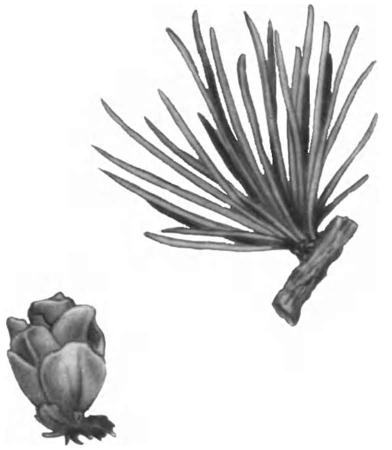
Homer V. Craig

Larch

A genus, *Larix*, of the pine family, with deciduous needles and short spurlike branches, which annually bear a crown of needles. The cones are small and persistent, varying by species in size, number, and form of the cone scales. The tamarack (*L. laricina*), also called hackmatack, is a native species, has erect, narrowly pyramidal habit, and grows in wet and moist soils in the northeastern United States, west to the Lake states, and across Canada to Alaska. The cones are $^1/_2$–$^3/_4$ in. (12–19 mm) long (see **illus.**). The tough resinous wood is durable in contact with the soil and is used for railroad ties, posts, sills, and boats. Other uses include the manufacture of excelsior, cabinet work, interior finish, and utility poles. *See* PINALES.

The western larch (*L. occidentalis*), the most important and largest of all the species, grows in the northwestern United States and southeastern British Columbia. The cones of this species are larger, 1–$1^1/_2$ in. (25–38 mm) long, with bracts growing out beyond the cone scales. The trunk is tall and erect, sometimes attaining a height of 200 ft (60 m) and a diameter of 6–7 ft (1.8–2.1 m). More than one-half of the annual production of larch comes from Montana, the remainder from Idaho, Washington, and Oregon.

The European larch (*L. decidua*) has cones about twice the size of those of the tamarack and 40–50 scales to a cone, whereas tamarack has only



Cone and needles of tamarack (*Larix laricina*).

12–15. The European larch does better in drier soil and is the species usually planted in parks and private grounds.

Golden larch (*Pseudolarix amabilis*), from China, is occasionally cultivated. Its leaves, also deciduous, are golden yellow in the fall. The cone scales fall off one by one, leaving the central axis of the cone on the tree. *See* FOREST AND FORESTRY; TREE.

Arthur H. Graves; Kenneth P. Davis

Large systems control theory

A branch of control theory concerned with large-scale systems. The three commonly accepted definitions of a large-scale system are based on notions of decomposition, complexity, and centrality. A system is sometimes considered to be large-scale if it can be partitioned or decomposed into small-scale subsystems. Another definition is that a system is large-scale if it is complex; that is, conventional techniques of modeling, analysis, control, design, and computation do not give reasonable solutions with reasonable effort. A third definition is based on the notion of centrality. Until the advent of large-scale systems, almost all control systems analysis and design procedures were limited to components and information grouped in one geographical location or center. Thus, by another definition, a system in which the concept of centrality fails is large-scale. This can be due to a lack of either centralized computing capability or a centralized information structure. Large-scale systems appear in such diversified fields as sociology, management, the economy, the environment, computer networks, power systems, transportation, aerospace, robotics, manufacturing, and navigation. Some examples of large-scale systems are the United States economy, the global telephone communication network, and the electric power generation system for the western United States.

Large-scale control systems have been an active research area for some time, but the field is still in its infancy. Nevertheless, it has set forth foundations for the systematic treatment of a class of very complex physical systems.

Since the 1950s, system theory has evolved from a semiheuristic discipline directed toward the design and analysis of electronic and aerospace systems containing few components into a very sophisticated theory capable of treating large, complex systems with myriad applications. Initially, system engineers attempted to cope with the increasing complexity through the development of sophisticated numerical techniques in order to apply classical theory to large systems. This approach, however, soon reached a point of diminishing returns, and it became apparent that new theoretical techniques would be necessary. Although many such techniques are still being developed, it is generally accepted that the key to the successful treatment of a large-scale system is to exploit fully its structural interconnection. This takes place in two ways: by the full use of sparse matrix techniques or by the decomposition of larger systems into smaller ones. *See* CONTROL SYSTEMS; SYSTEMS ENGINEERING.

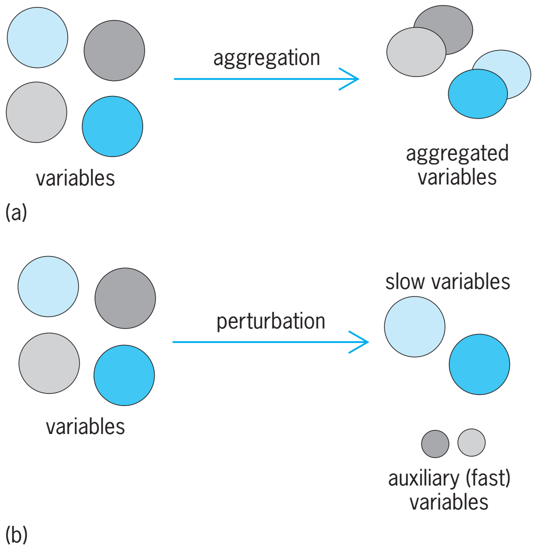


Fig. 1. Model reduction techniques. (*a*) Aggregation. (*b*) Perturbation.

Modeling and model reduction. The first step in any scientific or technological study of a system is to obtain a mathematical model. The major effort in large-scale system modeling has been reduction by aggregation and perturbation. An aggregate model of a system is described by a coarser set of variables. The underlying reason for aggregating a system model is to retain the key qualitative properties of the system, such as stability. Perturbation, on the other hand, is based on ignoring certain interactions of a system. Here again, however, the key properties must not be sacrificed for the sake of reduced computation. The common feature of both aggregation and perturbation is the reduction of system order. **Figure 1** depicts the way in which a fourth-order system is reduced to second order by both schemes. In aggregation, a combination (mixture) of the states (circles) is retained, whereas in perturbation the slow variables are retained and the fast ones are not. *See* PERTURBATION (MATHEMATICS).

In many applications of complex control systems, the accuracy of the model can come into question. To overcome uncertainties and inaccuracies in the model, a larger view of the complex system, encompassing the model, its past history, operational information, and so forth, should be addressed. This view may be designated collectively as knowledge. *See* MODEL THEORY.

Hierarchical control. A large-scale system, as defined above, can be described as a complex system composed of a number of constituents or smaller subsystems serving particular functions, sharing resources, and governed by interrelated goals and constraints. Although interaction among subsystems can take many forms, one of the most common is hierarchical, which appears to be natural in economic management, organizations, and complex industrial

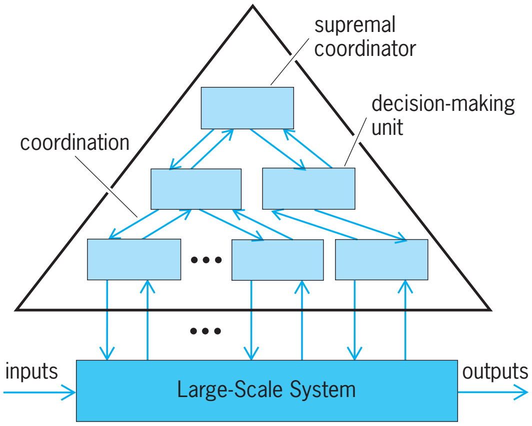


Fig. 2. **Pyramidal structure of a hierarchical large-scale system. Each row of boxes represents a hierarchical level. Downward vertical arrows represent control actions, and upward vertical arrows represent output measurements.**

systems. Within the hierarchical structure, the subsystems are positioned on different levels. A subsystem at a given level controls or coordinates the units on the level below it and is in turn controlled or coordinated by the unit immediately above.

Figure 2 shows a typical hierarchical (multilevel) system. The highest-level coordinator, sometimes called the supremal coordinator, can be thought of as the board of directors of a corporation, whereas other coordinators may be the president, vice president, and directors. Lower levels are occupied by the plant manager or shop managers; the large-scale system is the corporation itself. In spite of this seemingly natural representation of a hierarchical structure, its exact behavior is not well understood, mainly because relatively little quantitative work has been done on these large-scale systems.

There is no unique or universally accepted set of properties associated with hierarchical systems. However, the following are key properties: (1) a hierarchical system consists of decision-making components structured in a pyramidal shape (Fig. 2); (2) the system has an overall goal that may or may not be that of its components; (3) the various levels in the system exchange information (usually vertically) iteratively; and (4) as the level of hierarchy increases, the time horizon increases; that is, the lower-level components work faster than those at the higher level. *See* MULTILEVEL CONTROL THEORY.

Decentralized control. Most complex systems are characterized by a great multiplicity of measured outputs and inputs. For example, an electric power system has several control substations, each being responsible for the operation of a portion of the overall system. This situation arising in a control system design is often referred to as decentralization. The designer for such systems determines a structure for control which assigns system inputs to a given set of local controllers (stations) that observe only local system outputs. This approach, called decentralized control, attempts to avoid the difficulties associated with data gathering, storage requirements, computer program debugging, and geographical separation of system components.

In hierarchical control, a decomposition in system structure leads to computational efficiency. In decentralized control, simpler controller structures and computational efficiency result from a decomposition in the system's output information. The basic characteristic of any decentralized system is that the transfer of information from one group of sensors or actuators to others is quite restricted. For example, in the two-controller decentralized system of **Fig. 3**, only the output y_1 and external input v_1 are used to find the control u_1, and likewise the control u_2 is obtained through only the output y_2 and external input v_2.

The determination of control signals u_1, and u_2 based on the output signals y_1 and y_2, respectively, is nothing but two independent output feedback problems, which can be used for stabilization or pole-placement purposes. It is therefore clear that the decentralized control scheme is of feedback form, indicating that this method is very useful for complex linear systems. *See* CONTROL SYSTEM STABILITY; LINEAR SYSTEM ANALYSIS.

Attributes of large systems. Although there is no universally accepted set of attributes for large complex systems, the following may be a good starting point.

1. Large complex systems are often controlled by more than one controller or decision maker, involving decentralized computations.

2. The controllers have different but correlated information available to them, possibly at different times.

3. Large complex systems can also be controlled by local controllers at one level whose control actions are being coordinated at another level in a hierarchical (multilevel) structure.

4. Large complex systems are usually represented by imprecise aggregate models.

5. Controllers may operate in a group as a team or in a conflicting manner, with single- or multiple-objective or even conflicting-objective functions.

6. Large complex systems may be satisfactorily optimized by means of suboptimal or near-optimal controls, sometimes termed a satisfying strategy.

Autonomous control systems. The difficulties in developing a complex large-scale system model begin with its availability, and propagate through the

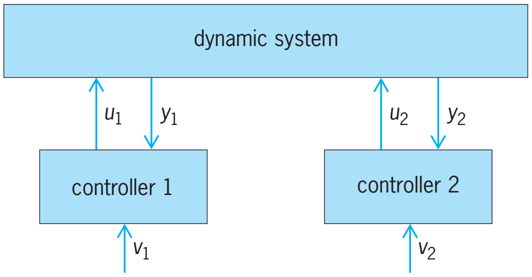


Fig. 3. **Two-controller decentralized system.**

simulation, analysis, and design of the system. Attempts have been made to introduce a new tool, artificial intelligence, into the modeling and control of systems—large or small in scale. Artificial intelligence can be defined as the science of automating intelligent behavior. The infusion of intelligence into the modeling, analysis, identification, design, and control of the system leads to a new generation of systems whose behavior is autonomous. An excellent example of an autonomous control system is NASA's *Pathfinder* spacecraft, which landed on Mars on July 4, 1997. Of course, such systems must be provided with human supervisory control and, in that sense, are not totally autonomous, but they do have a high degree of autonomy when functioning as intended by their human designers. *See* ARTIFICIAL INTELLIGENCE; SPACE PROBE.

Electric power systems. Another example of a large system is a regional power system. The system consists of many generating substations (subsystems) which are geographically separated but are linked through transmission lines. Here, the system has a number of local controllers, one at each substation (Fig. 3), which collectively work to ensure stability and reliability of the overall system. *See* CONTROL SYSTEMS; ELECTRIC POWER SYSTEMS.

Mohammad Jamshidi

Bibliography. Y. Y. Haimes (ed.), *Large Scale Systems*, North Holland, Amsterdam, 1982; M. Jamshidi, *Large-Scale Systems: Modeling and Control*, Elsevier, 1983; M. Jamshidi, *Large-Scale Systems Modeling, Control and Fuzzy Logic*, Prentice Hall, 1996.

Larmor precession

A precession in a magnetic field of the motion of charged particles or of particles possessing magnetic moments.

Charged particles. The Larmor theorem states that, for electrons moving in a single central field of force, the motion in a uniform magnetic field H is, to first order in H, the same as a possible motion in the absence of H except for the superposition of a common precession of angular frequency given by Eq. (1).

$$\omega_L = \frac{eH}{2mc} \tag{1}$$

Here e/c is the magnitude of the electronic charge in electromagnetic units, and m is the electronic mass. The frequency ω_L is called the Larmor frequency and is numerically equal to 2π times 1.40 MHz per oersted or 2π times 111 MHz per SI unit of magnetic field strength (ampere-turn per meter). *See* PRECESSION.

The Larmor theorem is derived in numerous texts. For the special case of an electron moving in a circular orbit of radius r about a fixed nucleus, with H applied normal to the plane of the orbit, the derivation is as follows: The centripetal force holding the electron in orbit must equal $m\omega^2 r$ and is the sum of the Coulomb force Ze^2/r^2 and the Lorentz force, $(e/c)\omega rH$. Therefore Eq. (2) is valid, where ω_0 is the

$$\begin{aligned}\omega &= \pm\left[\left(\frac{eH}{2mc}\right)^2 + \frac{Ze^2}{mr^3}\right]^{1/2} + \frac{eH}{2mc}\\ &= \pm({\omega_L}^2 + {\omega_0}^2)^{1/2} + \omega_L\end{aligned} \tag{2}$$

angular frequency in the absence of H. If $\omega_0 > \omega_L$ (bound electron, and first order in H) the approximate angular frequency is given by Eq. (3), which

$$\omega = \pm\omega_0 + \omega_L \tag{3}$$

is the Larmor theorem. For a free or unbound electron (no Coulomb force), the approximation breaks down, but direct solution of the equation involving $m\omega^2 r$ and the Lorentz force yields $\omega = eH/mc$. This is twice the Larmor frequency and is called the cyclotron frequency. *See* PARTICLE ACCELERATOR.

In stating the Larmor theorem, use was made of the phrase "a possible motion." if H is applied sufficiently slowly, it can be proved that the motion is the same as in the absence of H, except for the superposition of the Larmor precession. However, a sudden application of H may change, for example, a circular orbit into an elliptical one. For an important application of the Larmor theorem *see* DIAMAGNETISM; ELECTRON MOTION IN VACUUM.

Magnetic moments. According to elementary electromagnetic theory, a current loop of area A and of current I possesses a magnetic moment μ of magnitude IA and of direction normal to the loop. Thus an electron moving with a velocity v in a circular orbit of radius r, and hence with current $(-e/c)\cdot(v/2\pi r)$ in emu, has an orbital magnetic moment of magnitude as given by Eq. (4).

$$\mu = \frac{-e}{c}\frac{v}{2\pi r}(\pi r^2) = -\frac{evr}{2c} \tag{4}$$

The electron also has orbital angular momentum mvr, which by quantum theory must equal $\hbar J$, where J is an integer and $\hbar$ is Planck's constant h divided by 2π. The ratio of magnetic moment to angular momentum, Eq. (5), is called the magnetogyric (and

$$\gamma_J \equiv \frac{\mu}{\hbar J} = -\frac{e}{2mc} \tag{5}$$

often the gyromagnetic) factor γ_J. *See* ANGULAR MOMENTUM; GYROMAGNETIC RATIO; NONRELATIVISTIC QUANTUM THEORY.

In terms of the equivalent magnetic moment, Eq. (1) may be written in the form of Eq. (6). In

$$\omega_L = -\gamma_J H = -\frac{\mu}{\hbar J}H \tag{6}$$

this form the Larmor precession is exhibited by any magnetic moment μ including magnetic moments associated with spin angular momentum as well as those associated with orbital angular momentum. Equation (6) may also be derived from equating the time rate of change of angular momentum $(d/dt)(\hbar\mathbf{J})$

to the magnetic torque $\boldsymbol{\mu} \times \mathbf{H}$, as in Eq. (7). In

$$\frac{d(\hbar\mathbf{J})}{dt} = \mu \times \mathrm{H} = \gamma_J \hbar \mathbf{J} \times \mathbf{H} \qquad (7)$$

this form the Larmor precession applies to experiments in molecular beams, electron paramagnetic resonance (EPR), and nuclear magnetic resonance (NMR). *See* ELECTRON PARAMAGNETIC RESONANCE (EPR) SPECTROSCOPY; ELECTRON SPIN; MAGNETIC RESONANCE.

Rotating coordinate system. Let $(\partial/\partial t)$ represent differentiation with respect to a coordinate system rotating with angular velocity $\boldsymbol{\omega}$. Then differentiation with respect to a stationary observer (d/dt) is given by Eq. (8). Here $\mathbf{J}$ is measured by the stationary ob-

$$\frac{d\mathbf{J}}{dt} = \frac{\partial \mathbf{J}}{\partial t} + (\omega \times \mathrm{J}) \qquad (8)$$

server. This equation may be combined with Eq. (7) in the form of Eq. (9). Here $\mathbf{H}_r$ is the effective field in the rotating coordinate system as given by Eq. (10).

$$\frac{\partial(\hbar\mathbf{J})}{\partial t} = \gamma_J \hbar \mathbf{J} \times \left(\mathbf{H} + \frac{\omega}{\gamma_\mathrm{J}}\right) = \gamma_J \hbar \mathbf{J} \times \mathbf{H}_r \qquad (9)$$

$$\mathbf{H}_r = \mathbf{H} + \frac{\omega}{\gamma_\mathrm{J}} \qquad (10)$$

Therefore, in a frame which is rotating at the Larmor frequency, the effect of a constant field **H** is reduced to zero.

This result, which is an extension of Larmor's original theorem, also holds in quantum mechanics. It is the basis of simplified analyses of the effects of oscillating magnetic fields on particles with charges and magnetic moments.

Elihu Abrahams; Frederic Keffer

Bibliography. R. P. Feynman, *The Feynman Lectures on Physics*, 1964; H. Goldstein, *Classical Mechanics*, 2d ed., 1980; E. M. Purcell, *Electricity and Magnetism*, 2d ed., 1984; K. R. Symon, *Mechanics*, 3d ed., 1971.

Larnite

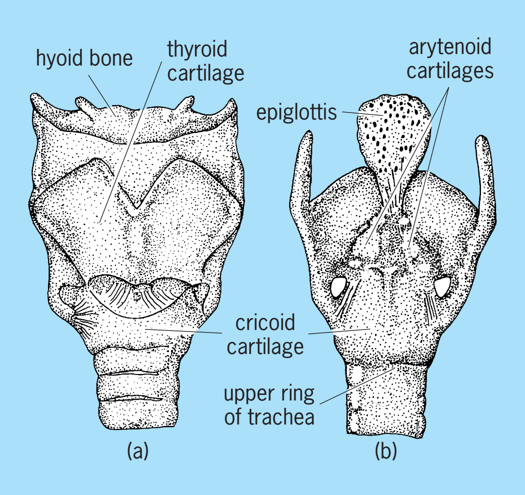
The alpha polymorph of calcium silicate (Ca_2SiO_4). Larnite is a mineral which crystallizes at high temperature. Its occurrences are practically confined to limestone or chalk zones in contact with semimolten balsalts. The presence of magnesium leads to the formation of other structurally related high-temperature minerals such as merwinite ($Ca_3Mg[SiO_4]_2$), bredigite ($Ca_7Mg[SiO_4]_4$), and alite ($Ca_3O[SiO_4]$). These phases are structurally very dense and have been postulated as possible components in the Earth's mantle. At least four polymorphs of Ca_2SiO_4 are known, and a few others are stabilized by impurities. They often occur in silicate slags and clinkers.

Larnite is stable between 1300 and 2600°F (720 and 1400°C) in the CaO-rich portion of the system $CaO–SiO_2$. At room temperature, larnite is metastable and inverts to its low-temperature polymorph calcio olivine through shock. This leads to fall, or distintegration of slags with time, and presents problems in the cement industry. The mineral is very rare, known from its type locality at Scawt Hill, County Antrim, Ireland, and from Crestmore, near Riverside, California. *See* SILICATE MINERALS.

Paul B. Moore

Larynx

The complex of cartilages and related structures at the opening of the trachea, or windpipe, into the pharynx, or throat.



Human laryngeal cartilages and ligaments. (*a*) Front view. (*b*) Back view. Epiglottis acts as lid to larynx.

The cartilages of the larynx (see **illus.**) are derivatives of the primitive gill-bar system. Their original function is to act as a protective sphincter at the entrance of the air passage into the throat, and, when this is the only function, the structure is quite simple, with only one or two pairs of cartilages.

In higher vertebrates the additional function of phonation has been acquired; it is almost limited to mammals since in birds the sound is produced in the syrinx lower in the trachea, but some frogs and reptiles can produce sounds with their larynges.

In humans and most other mammals, the signet-shaped cricoid cartilage forms the base of the larynx and rests upon the trachea. The thyroid cartilage, which forms the prominent Adam's apple ventrally, lies anterior to the cricoid. Dorsally there are paired pivoting cartilages, the arytenoids. Each is pyramid-shaped and acts as the movable posterior attachment for the vocal cords and the laryngeal muscles that regulate the cords. Two other small paired cartilages, the cuneiform and the corniculate, also lie dorsal to the thyroid cartilage. The epiglottis, a leaf-shaped elastic cartilage with its stem inserted into the thyroid notch, forms a lid to the larynx. *See* LARYNX DISORDERS; SPEECH; THYROID GLAND.

Thomas S. Parsons

Larynx disorders

Diseases of the larynx manifest themselves by hoarseness and by stridor, a form of noisy breathing caused by localized narrowing in the larynx or trachea.

Laryngitis is an inflammation of the mucous membrane of the larynx always associated with hoarseness. It frequently occurs with common colds and as a complication of other inflammatory diseases of the upper respiratory system. In diphtheria, the formation of a membrane of fibrin, leukocytes, destroyed tissue, and bacteria can cause severe respiratory difficulties, which may demand tracheotomy. The development of chronic laryngitis is favored by a chronic irritation such as that caused by smoking.

Hoarseness is also a manifestation of paralysis of the recurrent laryngeal nerve of the vagus, which is easily damaged upon surgical removal of a goiter.

Benign tumors of the larynx occur in the younger age group. A tumorlike formation of the vocal cord, known as singer's node, often accounts for the hoarseness of people who abuse the voice. Cancer of the larynx is not uncommon in humans, but is frequently of a slowly growing type and the outlook is often good. The highest incidence is in males over 60.

Ewald R. Weibel

Laser

A device that uses the principle of amplification of electromagnetic waves by stimulated emission of radiation, and operates in the infrared, visible, or ultraviolet region. The term laser is an acronym for light amplification by stimulated emission of radiation, or a light amplifier. However, just as an electronic amplifier can be made into an oscillator by feeding appropriately phased output back into the input, so the laser light amplifier can be made into a laser oscillator, which is really a light source. Laser oscillators are so much more common than laser amplifiers that the unmodified word "laser" has come to mean the oscillator, while the modifier "amplifier" is generally used when the oscillator is not intended. *See* AMPLIFIER; MASER; OSCILLATOR.

The process of stimulated emission can be described as follows: When atoms, ions, or molecules absorb energy, they can emit light spontaneously (as with an incandescent lamp) or they can be stimulated to emit by a light wave. This stimulated emission is the opposite of (stimulated) absorption, where unexcited matter is stimulated into an excited state by a light wave. If a collection of atoms is prepared (pumped) so that more are initially excited than unexcited (population inversion), an incident light wave will stimulate more emission than absorption, and there is net amplification of the incident light beam. This is the way the laser amplifier works.

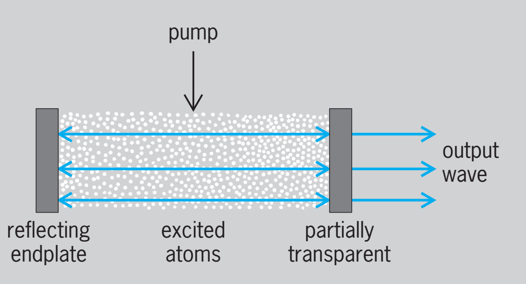


Fig. 1. Structure of a parallel-plate laser.

As with an electronic amplifier, the condition for oscillation is to feed back, in phase with the input, enough of the output to overcome the losses. A laser amplifier can be made into a laser oscillator by arranging suitable mirrors on either end of the amplifier. Together these are called the resonator. Thus the essential parts of a laser oscillator are an amplifying medium, a source of pump power, and a resonator. Radiation that is directed straight along the axis bounces back and forth between the mirrors and can remain in the resonator long enough to build up a strong oscillation. (Waves oriented in other directions soon pass off the edge of the mirrors and are lost before they are much amplified.) Radiation may be coupled out by making one mirror partially transparent so that part of the amplified light can emerge through it (**Fig. 1**). The output wave, like most of the waves being amplified between the mirrors, travels along the axis and is thus very nearly a plane wave. *See* OPTICAL PUMPING.

Comparison with other sources. In contrast to lasers, all conventional light sources are basically hot bodies which radiate by spontaneous emission. The electrons in the tungsten filament of an incandescent lamp are agitated by, and acquire excitation from, the high temperature of the filament. Once excited, they emit light in all directions and revert to a lower energy state. Similarly, in a gas lamp the electric current excites the atoms to high-energy quantum states, and they soon give up this excitation energy by radiating it as light. In all the above, spontaneous emission from each excited electron or atom takes place independently of emission from the others. Thus the overall wave produced by a conventional light source is a jumble of waves from the numerous individual atoms. The phase of the wave emitted by one atom has no relation to the phase emitted by any other atom, so that the overall phase of the light fluctuates randomly from moment to moment and place to place. The lack of correlation is called incoherence.

Hot bodies emit radiation, more or less equally in all directions, whose wavelength distribution is dictated by the Planck blackbody radiation curve. For example, the surface of the Sun radiates like a blackbody at a temperature of about 6000 K, and emits a total of 7 kW/cm^2, spread out over all wavelengths and directions. Light from gas lamps can be more monochromatic (wavelengths radiated are restricted by the quantized energies allowed in atoms), but radiation still occurs in all directions. In contrast, an ideal plane wave would have the same phase all across any wavefront, and the time fluctuations would be highly predictable (coherent). The output of the parallel-plate laser described above is very nearly such a plane wave and is therefore highly directional. This arises because in the laser oscillator atoms are stimulated to emit in phase with the stimulating wave, rather than independently, and the wave that builds up between the mirrors matches very closely the mirror surfaces. The output is powerful because atoms can be stimulated to emit much faster than they would spontaneously. It is highly monochromatic largely because stimulated emission is a resonance process that occurs most rapidly at the center of the range of wavelengths that would be emitted spontaneously. Since atoms are stimulated to emit in phase with the existing wave, the phase is preserved over many cycles, resulting in the high degree of time coherence of laser radiation.

As a result of this high degree of time coherence, combined with the excellent spatial coherence imposed by the mirror surfaces, lasers can emit vastly more power per unit area and per solid angle than conventional sources. Although such a comparison underscores the spectacular advance in this regard, the difference is not fundamental, but one of magnitude. A truly fundamental difference between laser light and conventional (thermal) light is in the photoelectron statistics when the light is detected. This particular aspect of laser light has yet to be widely utilized, but is crucial for the implementation of quantum computing concepts. *See* COHERENCE; HEAT RADIATION; RADIANCE.

Continuous-wave gas lasers. Perhaps the best-known gas laser is the neutral-atom helium-neon (HeNe) laser, which is an electric-discharge-excited laser involving the noble gases helium and neon. The lasing atom is neon. The wavelength of the transition most used is 632.8 nanometers; however, many helium-neon lasers operate at longer and shorter wavelengths including 3390, 1152, 612, 594, and 543 nm. Output powers are mostly around 1 milliwatt, although at 632.8 nm multimode powers approaching 100 mW can be obtained.

A useful gas laser for the near-ultraviolet region is the helium-cadmium (HeCd) laser, where lasing takes place from singly ionized cadmium. Wavelengths are 325 and 442 nm, with powers up to 150 mW.

The argon ion laser provides continuous-wave (CW) powers up to about 50 W, with principal wavelengths of 514.5 and 488 nm, and a number of weaker transitions at nearby wavelengths. The argon laser is often used to pump other lasers, most importantly tunable dye lasers and titanium: sapphire lasers. Argon lasers have the provision for selecting individual transition wavelengths for oscillation. Single-frequency oscillation can also be produced by inserting an etalon inside the laser resonator. For applications requiring continuous-wave power in the red, the krypton ion laser can provide continuous-wave lasing at 647.1 and 676.4 nm (as well as 521, 568, and other wavelengths), with powers somewhat less than those of the argon ion laser.

The carbon dioxide (CO_2) molecular laser has become the laser of choice for many industrial applications, such as cutting and welding. Oscillation can be obtained on more than 200 vibration-rotation transitions in the 8–18-micrometer wavelength region. The principal transitions are at 10.6 and 9.6 μm, with discrete tuning possible near both. Continuous-wave powers as large as hundreds of kilowatts are obtainable for industrial lasers, making it possible for a factory to use a central, very high power laser and distribute its output as needed. The efficiency of this laser exceeds 10%, which is very desirable for a high-power laser. Terawatts (10^{12} W) of peak power have been obtained in pulsed operation.

Short-pulsed gas lasers. Some lasers can be made to operate only in a pulsed mode. Examples of self-terminating gas lasers are the nitrogen laser (337 nm) and excimer lasers (150 to 400 nm). The nitrogen laser pulse duration is limited because the lower level becomes populated due to stimulated transitions from the upper lasing level, thus introducing absorption at the lasing wavelength. Peak powers as large as 1 MW are possible with pulse durations of 1–10 nanoseconds. Excimer lasers are self-terminating because lasing transitions tear apart the excimer molecules and time is required for fresh molecules to replace them. Other powerful lasers for the vacuum ultraviolet region (100 to 300 nm) use pulsed discharges in high-pressure molecular hydrogen or rare gases, such as xenon. These lasers provide nanosecond pulses with peak powers from tens of kilowatts to a few megawatts.

The nitrogen laser (which is suitable for home workshop laser construction) has proven useful for pumping tunable dye lasers, which lase most readily when pumped with short, intense pulses. Excimer lasers emit at even shorter wavelengths, down to 157 nm. They have proven useful in biological applications because their short wavelengths make possible selective severing of intermolecular bonds. Thus, very fine cutting can be accomplished without excessive heating. The short wavelengths of excimer lasers also allow for high-resolution patterning of photoresist in optical lithography. *See* LASER PHOTOBIOLOGY.

The copper vapor laser is a high-gain neutral-atom laser which operates (pulsed only) at two visible wavelengths, 510.6 and 578.2 nm. Typical pulse durations are 50 ns. A very large copper laser system was built for astronomy applications, which uses a high-repetition-rate laser (5000 pulses per second) followed by three laser amplifiers to achieve an average power of 250 W (peak power of 1 MW). Other applications include pumping large tunable dye lasers for isotope separation and precision machining and cutting, where the short-wavelength diffraction limit is 20 times smaller than for carbon dioxide lasers. *See* ADAPTIVE OPTICS; ISOTOPE SEPARATION.

Solid-state lasers. The term solid-state laser should logically cover all lasers other than gaseous or liquid. Nevertheless, current terminology will be followed, which treats semiconductor (diode) lasers separately from solid-state lasers because the physical mechanisms are somewhat different. With that reservation, virtually all solid-state lasers are optically pumped. Except for some brief remarks about the ruby laser,

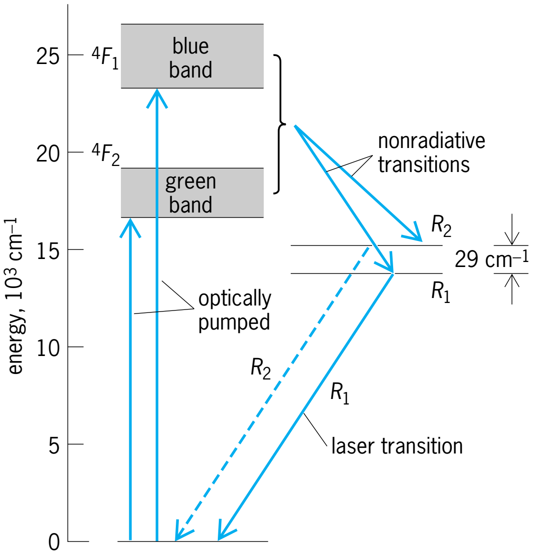


Fig. 2. Energy-level diagram of a ruby laser (three-level system).

the following discussion is limited to the main lasers in current use.

Ruby ($Cr{:}Al_2O_3$) laser. Historically, the first laser was a single crystal of synthetic ruby, which is aluminum oxide (Al_2O_3 or sapphire), doped with about 0.05% (by weight) chromium oxide (Cr_2O_3). The ruby laser (**Fig. 2**) actually comprises two laser systems with the upper laser levels, R1 and R2. Each is called a three-level system. The major pump bands are blue and green. Unless the ruby laser is pumped by another laser, only pulsed operation is possible. (The first ruby laser was pumped by a flashlamp.) The gain bandwidth is about 300 GHz at room temperature.

Rare-earth lasers and amplifiers. Three important rare-earth laser systems are neodymium:YAG, that is, yttrium aluminum garnet ($Y_3Al_5O_{12}$) doped with neodymium; neodymium:glass; and erbium:glass. Other rare earths and other host materials also find application. In contrast to ruby, the rare-earth ions form four-level systems (**Fig. 3**). This property results in a low pumping threshold for oscillation (about 1000 times lower than for ruby) because it is not necessary to invert the populations of the lower two levels, but only to maintain an inversion between levels 2 and 3. (Level 3 generally empties rapidly.)

1. *Neodymium:YAG laser.* Continuous-wave operation is possible with pumping from flash lamps or efficient diode lasers, typically at a wavelength of 808 nm. The dominant laser transition wavelength is 1064 nm. The gain bandwidth at room temperature is 60 GHz. The overall electrical-to-optical power conversion efficiency with diode laser pumping can exceed 20%. Power scaling of neodymium:YAG lasers is limited to a few kilowatts primarily due to thermal lensing. The laser radiation can be converted efficiently by frequency doubling or frequency tripling to yield green (532 nm) or blue (355 nm) light.

2. *Neodymium:glass laser.* In a typical glass host, neodymium has broader pump bands than in crystalline YAG, which enhance broadband pumping, but also result in a broader gain bandwidth (3000 GHz), which reduces the maximum gain. As a result, neodymium: glass lasers generally have a higher lasing threshold than neodymium:YAG lasers, but prove to be an excellent laser oscillator and amplifier particularly well suited for laser pulses with picosecond to nanosecond duration. Since glass has 40 times poorer thermal conductivity than YAG, it is difficult to remove heat. This severely limits the repetition rate for neodymium:glass lasers. On the positive side, large and uniform glass "logs" can be made by using superb glass technology (**Fig. 4**). Q-switched pulses with peak powers greater than 100 TW are possible.

3. *Erbium:glass laser.* High-bandwidth optical communication systems utilize optical carrier waves in the spectral region around 1550 nm and fiber-optic

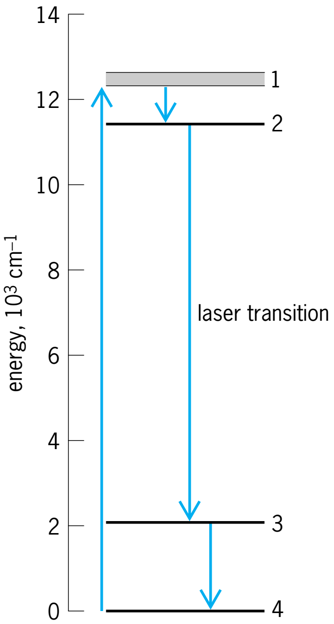


Fig. 3. Energy-level diagram of a rare-earth laser based on the Nd ion in YAG (four-level system). Broadband optical pumping is primarily to the broad, uppermost level. Alternatively, this level can be pumped by efficient semiconductor lasers.

Fig. 4. High-power laser amplifier stage, with a 30-cm (1-ft) aperture, using liquid-cooled slabs of neodymium:glass. (*Lawrence Livermore National Laboratory*)

cable with extremely low transmission losses at these wavelengths. Because of the broad gain band of erbium:glass in this wavelength region, this material is suited for 1550-nm glass lasers and the erbium-doped fiber amplifier (EDFA). Erbium:glass amplifiers and lasers are conveniently pumped by semiconductor diode lasers. The EDFA is a key element in optical communications. It is used to compensate for residual losses of long transmission lines, thereby replacing electronic repeaters which required transforming light to electricity and then back to light. Since signals at many carrier wavelengths (communication channels) can be amplified simultaneously in a broadband EDFA, these devices greatly improve the performance of optical communication systems.

Conventional rare-earth lasers and amplifiers use pencil-sized crystals or glass rods. Slab, thin-disk, and fiber lasers utilize special shapes and sizes of the amplifying medium to improve the laser performance. Slab, thin-disk, and fiber lasers provide much improved heat dissipation compared to traditional laser rods, making them suitable for high-power continuous-wave operation. Thin-disk lasers can generate output powers greater than 1 kW with optical pumping efficiencies up to 50% when pumped by high-power laser diodes.

Fiber lasers use guided-mode propagation for pump and signal light waves to create extremely robust laser structures that do not require any alignment after fabrication. Single-mode signal wave guiding in optical fibers results in diffraction-limited output beams from laser resonators that can be longer than 1 km. Fiber lasers can use all traditional rare-earth laser systems such as neodymium, erbium, or ytterbium. With single-mode output beams of more than 2 kW, continuous-wave power, and efficiencies over 80%, laser-diode-pumped ytterbium:glass fiber lasers are the most prominent high-power fiber lasers. Their emission wavelength of slightly above 1 μm is similar to neodymium:glass lasers. High-power laser diodes are used to pump the ytterbium ions at a wavelength of 980 nm. Erbium:glass fiber lasers that operate around 1550 nm can be used in fiber-optics-based communication and sensing systems that benefit from very long coherence lengths of fiber laser devices. *See* OPTICAL COMMUNICATIONS.

Titanium:sapphire ($Ti:Al_2O_3$) laser. The titanium-doped sapphire laser has become a very important solid-state laser, chiefly by virtue of its large gain bandwidth. Pumped by argon ion or diode lasers, it can be tuned from 660 to 1180 nm, more than 10 times the range of a single-dye laser such as rhodamine 6 G (570 to 610 nm). Commercial models are available with frequency doubling, tripling, and fourth-harmonic generation, respectively, covering the spectral ranges 350–450, 240–300, and 210–225 nm. Titanium:sapphire lasers can be mode-locked to produce femtosecond (10^{-15} s) pulses. Femtosecond titanium:sapphire oscillators typically generate 1 W of average power and, because of their large bandwidth and high saturation fluence, amplified titanium:sapphire laser systems can produce laser pulses with petawatt (10^{15} W) peak power.

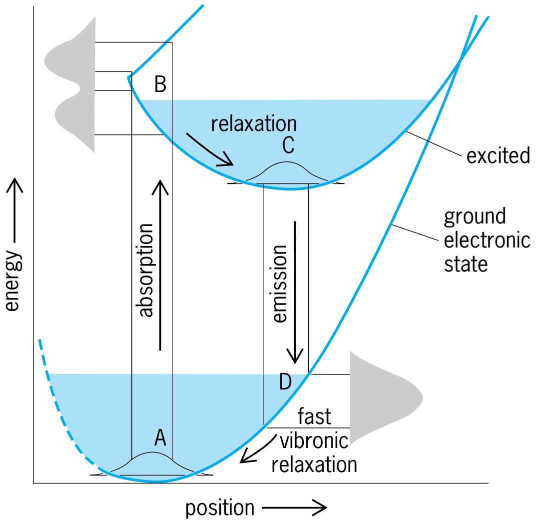


Fig. 5. Energy levels for a titanium-doped sapphire ($Ti:Al_2O_3$) laser. The energy of electronic states is plotted with respect to the position of the titanium ion. Shaded regions indicate continua of vibrational levels.

Since the structure of titanium:sapphire ($Ti:Al_2O_3$) seems similar to that of ruby ($Cr:Al_2O_3$), it might be supposed that titanium:sapphire is just a minor variation on the ruby laser, where one transition element, titanium, is substituted for another, chromium. It turns out that in this case the effect is not minor at all, for it results in a several-thousand-fold increase in the laser gain bandwidth. The reason has to do with the way that titanium ions orient themselves in the aluminum oxide lattice. When titanium ions in this lattice are in their excited states, the overall energy of the system can be lowered if the titanium ion is displaced with respect to the surrounding oxygen atoms (the Jahn-Teller effect). When the titanium ion moves to its new position, it excites lattice vibrations, or phonons (**Fig. 5**). Hence this laser is called a vibronic laser, as is the organic dye laser. When the titanium ion either absorbs or emits light, an atomic electron within the titanium ion reacts more quickly than the heavier titanium nucleus can move (the Franck-Condon principle). Thus, such electronic transitions can be represented as vertical lines in an energy-level diagram (Fig. 5). Gaussian-shaped curves (at points A and C in Fig. 5) can be used to represent the probability of finding the titanium ion at a particular position in the lowest vibrational state of the ground and excited electronic levels, respectively. When this laser is optically pumped in the blue-green region (transition A to B), the excited electron relaxes (to point C) by emitting phonons. Emission (from point C to D) can take place over a wide range of energy differences, which means a broad emission spectrum and wide gain bandwidth. In summary, the huge increase in gain bandwidth of the titanium:sapphire laser over that of the ruby laser is due to the Jahn-Teller effect. *See* FRANCK-CONDON PRINCIPLE; JAHN-TELLER EFFECT.

Semiconductor (diode) lasers. The semiconductor laser is the most important of all lasers, both by

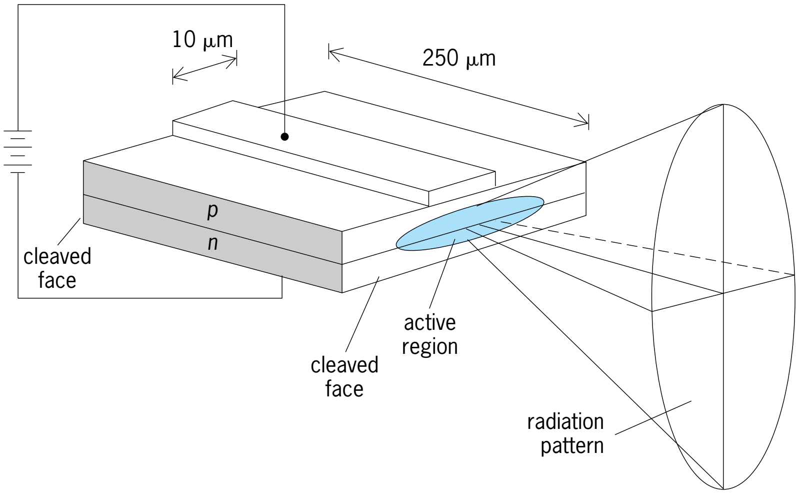


Fig. 6. Semiconductor *pn*-junction laser with horizontal cavity formed by cleaved end facets.

economic standards and by the degree of its applications. Its main features include rugged structure, small size (250 × 10 × 50 μm), high efficiency (sometimes exceeding 50%), direct pumping by low-power electric current (typically 15 milliamperes at 2 V, which makes it possible to drive it with transistor circuitry), ability to modulate its output by direct modulation of the pumping current at rates exceeding 20 GHz, compatibility of its output beam dimensions with those of optical fibers, feasibility of integrating it monolithically with other semiconductor optoelectronic devices to form integrated circuits, and a manufacturing technology that lends itself to mass production. *See* INTEGRATED OPTICS.

Most semiconductor lasers are based on III-V semiconductors. The laser can be a simple sandwich of *p*- and *n*-type material (**Fig. 6**) such as gallium arsenide (GaAs). The active region is at the junction of the *p* and *n* regions. Electrons and holes are injected through an electrical current into the active region from the *p* and *n* regions, respectively. Light is amplified by stimulating electron-hole recombination. A typical edge-emitting laser has a horizontal laser cavity that can be formed by cleaved-end facets of the chip, distributed Bragg reflectors (DBR), or a distributed feedback (DFB) scheme. *See* ELECTRON-HOLE RECOMBINATION; SEMICONDUCTOR; SEMICONDUCTOR DIODE.

It has been found that a sandwich of *p*-GaAlAs/GaAs/*n*-GaAlAs between the *p*-GaAs and *n*-GaAs layers better confines the light to the gain region and results in lower lasing thresholds. This device is called a double-heterostructure gallium arsenide–gallium aluminum arsenide (GaAs-GaAlAs) laser. GaAs-GaAlAs double-hetrostructure laser diodes reach power levels on the order of 100 mW, and lasing occurs around the wavelength of 800 nm.

Still lower lasing thresholds and higher efficiencies can be achieved using semiconductor nanostructures that provide structural confinement for electrons and holes in one, two, or three-dimensions. Most common are quantum-well structures where the carriers are confined through a potential barrier within nanometer-thin layers called quantum wells. The potential barrier is created through a change in material composition, that is, by including a third (or sometimes fourth) material such as aluminum, either in the barrier or in the well composition. The lowest lasing thresholds have been achieved in quantum-dot lasers, where the carriers are confined in all directions. Such three-dimensional confinement can be obtained through lithographic patterning or special growth modes leading to nanometer-sized semiconductor clusters. *See* ARTIFICIALLY LAYERED STRUCTURES; QUANTIZED ELECTRONIC STRUCTURE (QUEST); SEMICONDUCTOR HETEROSTRUCTURES.

Semiconductor laser diodes can emit at wavelengths from ultraviolet to the infrared with the use of a large variety of materials. Active semiconductor laser structures can include ternary as well as quaternary material compositions. The shortest wavelength ultraviolet diodes use gallium nitride (GaN)–based structures. Indium phosphide (InP)–based compositions are the material of choice at the important range of wavelengths in the optical communications window around 1550 nm. Further in the infrared, even longer wavelengths can be reached with quantum cascade lasers. In quantum cascade lasers, electrons cascade down an energy staircase between neighboring quantum wells, emitting a long-wavelength photon at each step. This is in contrast to all other laser diodes that have stimulating transitions between electron states in conduction bands and hole states in valence bands.

Broad-area semiconductor diode stripes can be fabricated to boost the output power of laser diodes. Lasers with stripes only a few micrometers wide can generate powers up to 100 mW, a 50-μm stripe can generate 500 mW, and more than 1-cm-wide stripes can generate more than 10 W. Monolithic arrays containing dozens of stripes can be manufactured. Power scaling through increasing stripe width, however, results in degradation of the optical beam quality. An alternative approach to generating high-power, high-quality beams is the master oscillator power amplifier (MOPA) structure. Here the high-quality output beam of a master oscillator is coupled into an optical amplifier which retains a virtually unchanged beam quality. Master oscillators and wide-area power amplifiers can be monolithically integrated into the same semiconductor chip, generating several watts of output.

Configurations different from traditional (horizontal resonator) laser diodes (Fig. 6) have been developed in the form of vertically emitting semiconductor lasers. Here the resonator is perpendicular to the plane of the active region, and the laser resonator is formed by mirrors on the top and bottom of the active region. If both mirrors are integrated directly on the chip, such a structure is called a vertical cavity surface-emitting laser (VCSEL), while a laser with one external output coupler is called a vertical external-cavity surface-emitting laser (VECSEL). In both configurations the light passes through only a short length of active medium (1 μm or less), resulting in low optical gain per cavity round-trip. Therefore, vertical-cavity lasers require mirrors with relatively high reflectivity. Although VCSELs typically have low continuous-wave output powers (a few milliwatts), their beams are more symmetrical than those of edge-emitting diodes. Furthermore, a VCSEL chip can be fabricated as a two-dimensional array of independently addressable laser elements. Although the more complex VECSEL requires one adjustable mirror and is not a completely monolithic laser device, it has the symmetrical beam shape of a surface emitter, it can generate tens of watts of power using a rather large area of the semiconductor chip, and it can provide wavelength tunability.

Low cost, small size, high power, and high efficiency have made semiconductor laser diodes useful in a wide range of applications, including compact-disk players, fiber-optic communications, optical data storage, laser printing, materials processing, and pumping of other (mostly solid-state) lasers. *See* COMPACT DISK; COMPUTER PERIPHERAL DEVICES; OPTICAL FIBERS; OPTICAL RECORDING.

Monochromaticity. When lasers were first developed, they were widely noted for their extreme monochromaticity. They provided far more optical power per spectral range (as well as per angular range) than was previously possible. It has since proven useful to relate laser frequencies to the international time standard (defined by an energy-level difference in the cesium atom), and this was done so precisely, through the use of optical heterodyne techniques, that the standard of length was redefined in such a way that the speed of light is fixed. (One meter is the distance that light travels in vacuum during 1/299,792,458 s.) In addition, extremely stable and monochromatic lasers have been developed, which can be used, for example, for optical communication between remote and moving frames, such as the Moon and the Earth. *See* FREQUENCY MEASUREMENT; HETERODYNE PRINCIPLE; LASER SPECTROSCOPY; LIGHT.

Tunable lasers. Having achieved lasers whose frequencies can be monochromatic, stable, and absolute (traceable to the time standard), the next goal is tunability. Most lasers allow modest tuning over the gain bandwidth of their amplifying medium. However, the laser most widely used for wide tunability has been the (liquid) dye laser. This laser must be optically pumped either by a flash lamp or by another laser, such as the argon ion laser. Considerable engineering has gone into the development of systems to rapidly flow the dye and to provide wavelength tunability. About 20 different dyes are required to cover the region from 270 to 1000 nm.

The dye laser is being challenged by other tunable lasers which do not require flowing dyes. In addition to the titanium:sapphire laser (described above), there is the narrow-band tunable external-cavity diode laser that can be configured as an edge emitter or as a surface emitter (see VECSEL described above). Since the tuning range of diode lasers is much smaller, separate lasers are needed to cover the range from 670 to 1500 nm (in steps of about 10 nm, with linewidths of about 1 MHz). However, compared to the dye laser, the external-cavity diode laser system is much simpler, more compact, and more efficient. Other broadly tunable lasers that are commercially available are the parametric oscillator and the color-center laser (for the region 0.8–4-μm).

Free-electron lasers. The purpose of the free-electron laser is to convert the kinetic energy in an electron beam to electromagnetic radiation. Since it is relatively simple to generate electron beams with peak powers of 10^{10} W, the free-electron laser has the potential for providing high optical power; and since there are no prescribed energy levels, as in the conventional laser, the free-electron laser can operate over a broad spectral range.

To accomplish an energy transfer from electrons to light over an extended distance, it is necessary that the phase velocity of the light wave equals the velocity of the current in the electron beam. This presents a problem, because in vacuum, light waves propagate at the velocity of light c, whereas the electron velocity is less than c. At microwave frequencies, this problem is solved by not having the wave propagate through vacuum but through a waveguide with conducting boundaries that vary periodically in the direction of wave motion. To match this boundary condition requires that a component of the wave propagates with a phase velocity less than c, and it is this component that can absorb energy from an electron beam. A microwave device based upon this principle is known as a traveling-wave tube. Such an approach does not work very well at infrared and optical wavelengths, because the dimensions of the periodic boundary are less than a wavelength, which requires submicrometer tolerances over meter lengths; and because the electromagnetic field decays exponentially from the periodic boundary in the direction normal to the direction of propagation, which limits the transverse interaction area to submicrometer dimensions. *See* TRAVELING-WAVE TUBE; WAVEGUIDE.

The free-electron laser does provide a solution for short wavelengths. Rather than having periodic boundaries for the wave, the electron beam is caused to move in a periodic fashion, which is usually accomplished by passing the beam through a spatially periodic magnetic field generated by an undulator magnet array (**Fig. 7**). Such a beam has a current component with a velocity greater than the electron velocity, and it is this component that delivers energy to the electromagnetic wave. Submicrometer fabrication tolerances are not needed, and the transverse interaction area can be large. Free-electron lasers have operated over the spectral range from microwave to ultraviolet. The far-infrared portion of the spectrum is a promising region of application since there are presently few alternative sources, free-electron lasers operate well in this region, and there are many interesting problems in physics and chemistry to be investigated. *See* SYNCHROTRON RADIATION.

High-power and short-pulse lasers. For high-power applications, it is interesting to compare gas lasers with solid-state lasers. Because they are condensed

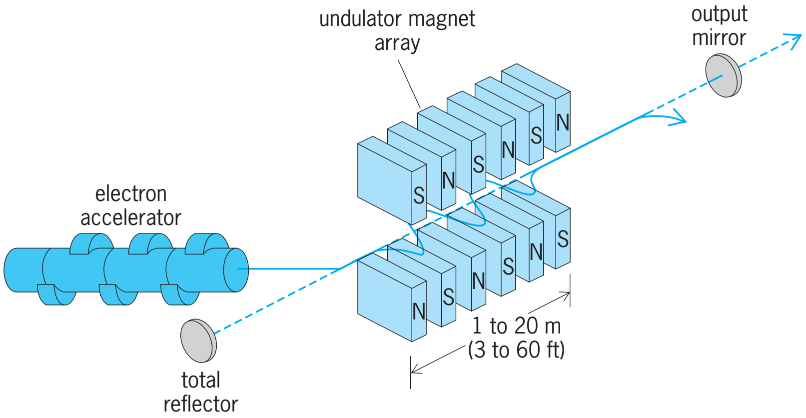


Fig. 7. Free-electron laser. Electrons vibrating in the undulator field amplify the optical beam.

matter, solid-state lasers can have much greater stored energy than gas lasers. However, gas lasers can partially compensate for this through mass transport, that is, replacing used atoms with freshly pumped ones by flowing the gas. The limiting factor in terms of power for gas lasers is the flow rate, which can be supersonic. The limiting factor for solid-state lasers is spontaneous emission lifetime, because flow is not possible, so that stored energy is limited by the time available to pump the same atoms.

For high peak power (pulsed operation), solid-state lasers excel. However, as the repetition rate is increased to increase the average power, the limitation becomes one of laser material damage. For solid-state lasers, the limitation is the heat dissipation rate, while for gas lasers the limitation is the mass transport rate. Gas lasers have the special advantage that if damage occurs, it is self-healing.

Some lasers, notably carbon dioxide and carbon monoxide (CO) lasers, which can emit continuously at modest powers, can also be used to generate intense pulses of microsecond duration when operated at high pressures. At high pressures it is difficult to obtain a uniform electric discharge across the region of lasing. One solution to this problem is the TEA (transverse-electric, atmospheric-pressure) laser, where the electric discharge takes place across the lasing column at pressures of about 1 atm. Still higher powers at higher pressures are obtainable through the ionizing effect of intense bursts of fast electrons from a small accelerator, known as an E-beam.

Short-pulsed operation is achievable from most lasers by a variety of special techniques, including Q-switching and mode-locking. Q-switching involves changing the loss, or Q, of the laser resonator, for example, by effectively misaligning mirrors using a shutter or an optical modulator. The amplifying medium thus becomes excited beyond the ordinary need for lasing, so that additional energy is stored in the medium. At this time, the loss is abruptly removed and the stored energy is released in a giant pulse (1–100 MW, typically lasting 1–30 ns). The Q-switch method works only with lasers whose upper states have long lifetimes. Still shorter and higher-power pulses can be generated by mode-locking or mode generation techniques, sometimes used in conjunction with Q-switching. Modes of a laser oscillator can be synchronized so that the peaks of their light waves occur simultaneously at some instant. The result is a short, intense pulse which quickly ends when the waves of different frequencies get out of step. Mode-locked lasers have generated pulses shorter than 10 femtosecond. Further pulse-shortening techniques have been developed to reach the attosecond (10^{-18} s) regime. *See* OPTICAL PULSES; Q (ELECTRICITY).

The beam may be further intensified by additional stages of amplification, provided the beam diameter is increased by some means such as a telescope, so as to expand (dilute) the beam and thereby prevent damage to laser materials and optics. Sometimes, for solid-state lasers such as neodymium:glass, the amplifying medium is divided into flat slabs separated by cooling liquid (Fig. 4). Here the open faces of the light-amplifying slabs present a large area to receive pumping light and liquid cooling.

Development of very large multistage lasers has been undertaken for research on thermonuclear fusion. In a particularly large laser of this type, the National Ignition Facility at Lawrence Livermore National Laboratory, a single neodymium:glass laser oscillator was designed to drive 192 neodymium:glass amplifier chains made up of together 3072 glass slabs. Each beam will deliver infrared pulses of 10-ns duration and more than 20-kJ energy. After frequency tripling, the combined power of about 500 TW (5×10^{12} W) will be focused onto a millimeter-sized metal target cylinder that contains a deuterium-tritium mixture. The experiment is designed to heat and compress the heavy hydrogen mixture until fusion between hydrogen nuclei is ignited to produce helium, releasing large amounts of nuclear energy. Ultimately this type of controlled laser fusion may become an important source of thermal, electrical, and chemical energy. *See* NUCLEAR FUSION.

Stephen F. Jacobs; Arthur L. Schawlow;
Richard Pantell; Axel Schülzgen;
Nasser Peyghambarian

Bibliography. C. C. Davis, *Lasers and Electro-Optics*, 1996; P. W. Milonni and J. H. Eberly, *Lasers*, 1988; B. E. A. Saleh and M. C. Teich, *Fundamentals of Photonics*, 1991; W. T. Silfvast, *Laser Fundamentals*, 2d ed., 2004; J. T. Verdeyen, *Laser Electronics*, 3d ed., 1995; A. Yariv, and P. Yeh, *Photonics: Optical Electronics in Modern Communications*, 6th ed., 2006.

Laser alloying

A material processing method which utilizes the high power density available from focused laser sources to melt metal coatings and a portion of the underlying substrate. Since the melting occurs in a very short time and only at the surface, the bulk of the material remains cool, thus serving as an intimate heat sink. Large temperature gradients exist across the boundary between the melted surface region and the underlying solid substrate. The result is rapid self-quenching and resolidification.

Transitions. The sequence of schematic cross sections in **Fig. 1** illustrates the transitions occurring during and following an individual laser exposure. In Fig. 1*a*, the metal substrate (B) coated with a thin metal film (A) is irradiated with a laser pulse. A fraction of the incident laser light is absorbed by free carriers within the electromagnetic skin depth of 10^{-7} to 10^{-6} in. (10^{-6} to 10^{-5} cm). For metal surfaces and most laser wavelengths, a significant fraction of the incident light will be specularly or diffusely scattered away (the reflectance process is shown in Fig. 1). The absorbed energy is "instantaneously" (10^{-12} s) transferred to the lattice. The near-surface region very rapidly reaches the melting point, and a liquid-solid interface starts to move through the film (Fig. 1*b*). In Fig. 1*c*, the liquid-solid interface has swept through the original thin-film–substrate interface. Interdiffusion of the film and substrate elements starts. The laser pulse is nearly terminated, and the surface has remained below the vaporization temperature. In Fig. 1*d*, the maximum melt depth has been reached, and interdiffusion continues. The resolidification interface velocity is momentarily zero and then rapidly increases. In Fig. 1*e*, the resolidification interface has moved approximately halfway back to the surface from the melt depth. Interdiffusion in the liquid continues, but the resolidified metal behind the liquid-solid interface cools so rapidly that solid-state diffusion may be neglected. In Fig. 1*f*, the material is completely resolidified, and a "surface alloy" of A in B has been produced. What makes laser surface alloying both attractive and interesting is the wide variety of chemical and microstructural states that can be retained because of the rapid quench from the liquid phase. These include chemical profiles where the "alloyed" element A is highly concentrated near the atomic surface and decreases in concentration over shallow depths (hundreds of nanometers), and uniform profiles where the concentration of A in B is the same throughout the entire melted region. The types of microstructures observed include extended solid solutions (the concentration of A in B greatly exceeds equilibrium values), metastable crystalline phases (high-temperature phases retained because of the rapid return to room temperature), and metallic glasses.

Surface alloying. In general, the output of a laser source must be focused in order to achieve sufficient energy density to induce melting of a metal surface. This focused spot is characterized by its optical spot size. The size of the melted spot (effective spot size) resulting from the laser exposure will generally be smaller than this optical spot size. For a continuous-wave laser source, a melt strip will be produced as the focused beam and metal surface are moved relative to one another. The dwell time, that is, the time that the continuous-wave laser beam irradiates a particular surface point, will strongly influence the depth of the melting. As is schematically shown in **Fig. 2***a*, individual melt strips must be partially

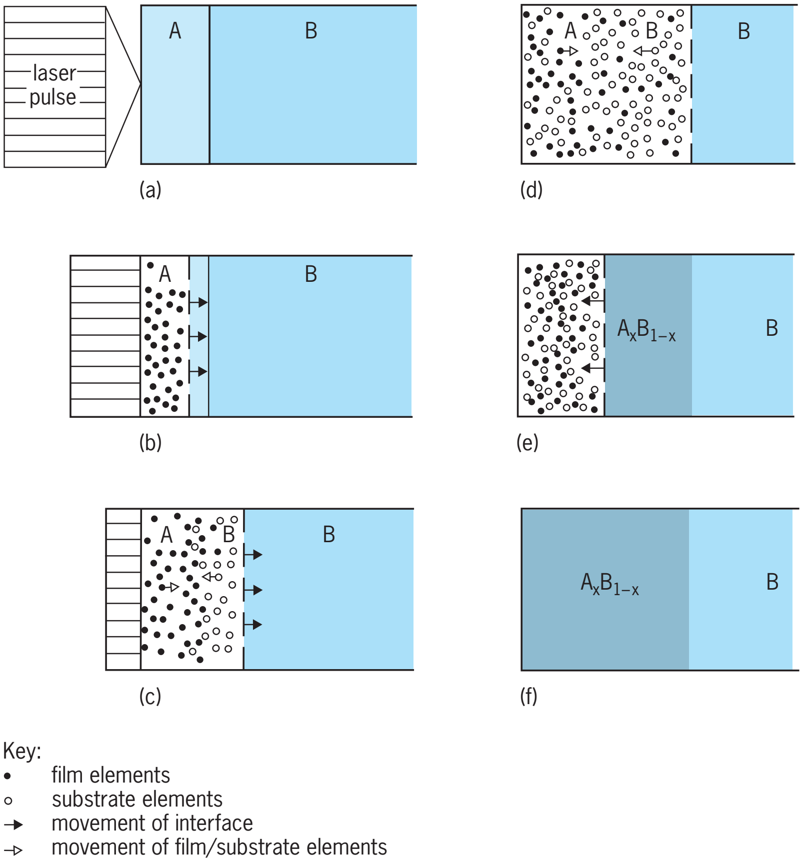


Fig. 1. **Sequence of schematic cross sections for laser alloying, with time increasing from *a* to *f*. Ax_xB_{1-x} represents surface alloy (with composition fixed by *x*) of film elements A in substrate elements B.**

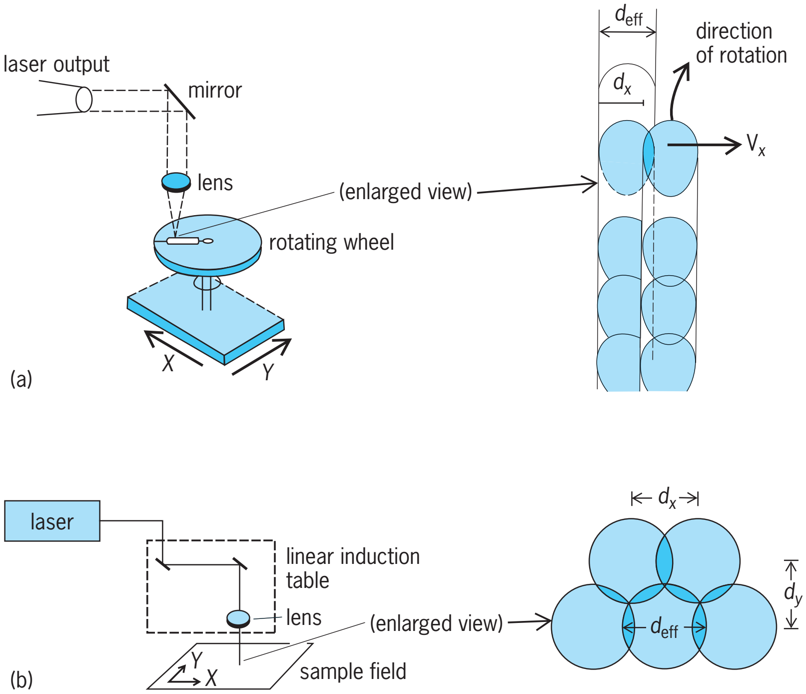


Fig. 2. **Schematic drawings and enlarged views of material processing for (*a*) continuous and (*b*) pulsed or Q-switched laser system. d_{eff} = diameter of melt spot or melt strip; V_x = velocity of sample movement in *x* direction; d_x, d_y = displacement between subsequent laser melt strips or spots resulting from relative motion of laser and sample.**

overlapped to produce area coverage. The effective spot size, degree of overlap, and relative scanning speed will thus determine the area per unit time which can be produced.

For a pulsed or Q-switched laser (high-power, short-pulse laser) source, there also is an optical and effective spot size, and the "pulse length" is the time of exposure. The laser pulses are emitted in a train of pulses characterized by a repetition rate. As is schematically shown in Fig. 2*b*, this train of pulses is raster-scanned across the metal surface to produce area coverage. For these lasers, the effective spot size, degree of overlap, and repetition rate determine the area per unit time which can be produced.

For all the laser sources (continuous-wave pulsed, and Q-switched), the exposure time (dwell time or pulse length) strongly influences the depth that will be melted. Longer exposure times result in deeper melting. Since deeper melting means a longer total time in the molten state, that means more time available for diffusion of the one or more alloying elements into the molten portion of the substrate. Deeper melting and longer melt times therefore result in more dilute surface alloys, while shallow melting and shorter melt times result in more concentrated surface alloys. It is also evident that in some instances convection, surface tension, and plasma effects can enhance the mixing within the liquid state and drive the melt toward homogenization.

In making laser alloys, many other processing variables need to be considered. In addition to the exposure time just discussed, these include the laser power, the thickness of the film put down prior to laser melting, and in some instances the nature of the gaseous ambient during the laser processing. The processing variables are interrelated, and one variable cannot be freely changed without affecting another. Another consideration is that laser alloying is a liquid state–rapid quenching phenomenon. The near-surface region must be melted and yet vaporization avoided. Different minimum and maximum energy densities are thus defined for each laser exposure time. In addition to these processing constraints, there are certain properties of matter which strongly influence whether or not certain element combinations may be laser-alloyed. For example, it is not possible to laser-alloy a low melting point–high vapor pressure element like zinc (Zn) into a high melting point–low vapor pressure metal substrate such as tungsten (W). The zinc would vaporize before the underlying tungsten could be melted. Since liquid-state intermixing is required, suitable systems must exhibit miscibility in the molten state. Binary systems like silver-nickel (Ag-Ni), iron-lead (Fe-Pb), copper-molybdenum (Cu-Mo), and aluminum-bismuth (Al-Bi) have miscibility gaps in the liquid state spanning nearly all compositions. Such systems cannot be laser-alloyed.

Advantages. Surface alloys, particularly laser surface alloys, have advantages that include tailoring surface properties, conservation of materials, and creation of new metal surfaces.

Surface alloying allows alteration of the metal surface in order to achieve characteristics best suited to the service environment. For example, in an ordinary saw blade the material requirements for the cutting teeth are very different from those of the length of the blade. In many applications requiring corrosion or oxidation resistance, that resistance is needed only on the external surface of the material. It is possible to design the bulk of the material with characteristics most suitable for structural needs or ease of fabrication and to tailor the metal surface for interface requirements. Material conservation is another reason often cited for considering surface alloying. In the case of stainless steels and superalloys, the elements added to impart the special properties are often strategic elements for which there are sparse domestic supplies. For example, chromium (Cr) and nickel (Ni) may not be necessary on the inside of knives, forks, and spoons. Further, precious metals are expensive and in relatively short supply. Thus surface alloying can be very cost-effective.

These first two advantages—tailoring for surface properties and material conservation—are generally important considerations in almost any coating technology. However, laser alloying differs from coating technologies in that the near-surface region is a continuous extension of the interior of the metal. There is no interface between bulk and "coating"; the laser alloy is a mixture of bulk and surface elements. Also, problems in regard to porosity and adherence do not exist. Synergistic benefits may be derived which are not realized by a coating alone. Laser alloying involves very large temperature gradients and quenching from the liquid state. In this way it resembles other rapid-solidification technologies. The thermodynamic constraints which limit the conventional metallurgist do not necessarily apply. It is anticipated that laser alloying will produce many new alloys which could not have been produced by conventional methods. *See* ALLOY; LASER; METAL COATINGS.

Clifton W. Draper

Bibliography. C. W. Draper, Laser surface alloying: A bibliography, *Appl. Opt.,*20(18):3093–3096, September 1981; C. W. Draper, Laser surface alloying: The state of the art, *J. Met.,*34(6):24–32, June 1982; D. S. Gnanamuthu, Laser surface treatment, *Opt. Eng.,* 19(5):783–792, 1980; D. K. Sood, Metastable surface alloys produced by ion implantation, laser of electron beam treatment, *Radiat. Effects,* 63(1):L141–167, January 1982.

Laser cooling

Reducing the thermal motion of atoms with the force exerted by a laser beam. Typically, such cooling is used to reduce the temperature of a gas of atoms, or the velocity spread of atoms in an atomic beam.

Light affects atomic motion when the atoms absorb or emit photons, the particles or quanta that make up light. Photons carry momentum $p = h/\lambda$, where h is Planck's constant and λ is the light's wavelength. By conservation of momentum, when an atom absorbs or emits a photon, the atom's momentum must change by an amount equal to the photon momentum. Each absorption or emission thus gives the atom a tiny kick, changing its velocity. For most atoms this change is only a few millimeters to a few centimeters per second, while atoms in a gas at room temperature have velocities of a few hundred to a few thousand meters per second. Nevertheless, repeated absorption and emission of photons can have a significant effect on even hot atomic gases or beams. *See* CONSERVATION OF MOMENTUM; LIGHT; MOMENTUM; PHOTON.

Doppler cooling. The keys to using such repeated kicks to reduce the random, thermal motion of a gas of atoms are the monochromatic nature of laser light, the selectivity of absorption of light by atoms, and the Doppler effect. Light is an oscillating electromagnetic wave whose frequency of oscillation determines its color. The energy of each photon is $E = h\nu$, where ν is the frequency. Laser light can have nearly a single frequency or color, so that all the photons have almost identical energies. Atoms absorb only photons whose energy is equal, within a small range, to the difference in energy between two of its quantum states or energy levels. For sodium atoms this resonance frequency is $\nu_0 \equiv 5 \times 10^{14}$ Hz (wavelength $\lambda \equiv 589$ nanometers), but the absorption is efficient only over a range $\Delta\nu = 10^7$ Hz. Moving atoms, however, experience a Doppler shift so that, depending on their speed and whether they are moving along the direction of the laser beam or against it, the light appears to room-temperature atoms to have a frequency shifted up or down by a hundred or more times the natural absorption width $\Delta\nu$. *See* DOPPLER EFFECT; LASER.

If the frequency ν of the laser is tuned to be slightly lower than ν_0, those atoms moving against the laser beam see the laser upshifted, closer to ν_0. These atoms are more likely to absorb photons, receive kicks opposite to the direction of their velocity, and slow down. After absorbing a photon, the atoms are in an excited state and return to the original state by spontaneously emitting a photon. Such photons are radiated in random directions, so the effect of their kicks averages to zero. For atoms held in a trap, as ions generally are, any trapped atom will at some time be traveling against the laser beam and be cooled. Laser cooling was first demonstrated in 1978 with such trapped ions. For free atoms, another, similarly tuned laser beam is added, aimed in the opposite sense, to cool those atoms moving in the opposite direction. More generally, one uses three pairs of mutually perpendicular, counterpropagating laser beams, all tuned below ν_0. Then, no matter the direction of an atom's velocity, there are one or more laser beams that oppose the velocity and slow the atom. Because atoms are slowed as if by friction in a viscous fluid, this configuration of laser beams is called optical molasses.

Cooling limit. Laser cooling's viscosity does not stop atoms. The discrete nature of the kicks imparted by the photons gives the atoms a random motion with a temperature as low as $h\Delta\nu/(2k_B)$, where k_B is Boltzmann's constant. For the example of sodium, this temperature is 240 microkelvins. Surprisingly, in 1988 much lower temperatures were observed in sodium and other atoms. The lower temperatures (as low as 1 μK for cesium atoms) were eventually explained as resulting from additional cooling processes that depend on the atom having multiple energy levels. Such temperatures, the lowest steady-state kinetic temperatures ever achieved for any substance in three dimensions, are close to the recoil limit, or the kinetic energy imparted to an atom from the kick of a single photon. While this might seem to be a fundamental limit, techniques have been developed that go even further, by arranging that when, by chance, an atom achieves a very low velocity it stops absorbing light and remains cold. In this way, atoms have been cooled below the recoil limit.

Applications. Improving atomic clocks, where the thermal motion of atoms reduces the precision and accuracy, was a major motivation to developing laser cooling. Clocks using laser-cooled trapped ions or free neutral atoms rival the performance of the best conventional atomic clocks. Laser cooling is also used in atom optics, where well-collimated, monoenergetic atomic beams are more easily and effectively manipulated. In addition, laser cooling has been used to study collisions between very slow atoms. *See* ATOMIC CLOCK; SCATTERING EXPERIMENTS (ATOMS AND MOLECULES).

Laser cooling is intimately connected with trapping of atoms, because atoms must often be slowed down before they can be held in a trap and because atoms must often be trapped in order to observe laser cooling or its effects. Such effects include cold, trapped ions arranging themselves into a crystal because of the electric repulsion between the charged ions. Neutral atoms can become arrayed on an optical lattice of tiny traps formed by interference between the laser beams used to cool them. In both cases, the spacing between atoms is thousands of times larger than the spacing in solid crystals. Another effect is Bose-Einstein condensation, wherein a gas of atoms whose de Broglie wavelength is comparable to the spacing between atoms has a transition to a state where a significant fraction of the atoms are in the lowest kinetic energy state possible. This phenomenon was first observed for a dilute atomic gas in 1995. *See* BOSE-EINSTEIN STATISTICS; PARTICLE TRAP; QUANTUM MECHANICS.

William D. Phillips

Bibliography. E. Arimondo, W. D. Phillips, and F. Strumia (eds.), *Laser Manipulation of Atoms and Ions*, 1992; S. Chu, Laser trapping of neutral particles, *Sci. Amer.*, 266(2):71–76, February 1992; S. Chu and C. Wieman (eds.), Laser cooling and trapping of atoms (feature issue), *J. Opt. Soc. Amer.*, B11:2019–2278, 1989; C. N. Cohen-Tannoudji and W. D. Phillips, New mechanisms for laser cooling, *Phys. Today*, 43(10):33–40, October 1990; D. J. Wineland and W. M. Itano, Laser cooling, *Phys. Today*, 40(6):34–40, June 1987.

Laser photobiology

The interaction of laser light with biological molecules, and the applications to biology and medicine. The spatial coherence of laser emissions makes it possible to focus beams in small volumes. Consequently, high-power-density irradiations can be obtained, especially with lasers whose energy is delivered in a few nanoseconds or even picoseconds. These two characteristics, combined with the great variety of wavelengths available, have opened up many possibilities of applications to biology and medicine.

Physical aspects. In the visible range, absorption of laser light by biological material is caused only by pigments such as chlorophyll, hemoglobin, cytochromes, or melanin, while ultraviolet laser light is absorbed by nucleic acids and proteins. At low intensity, the absorption of photons produces photochemical effects similar to those obtained with ordinary light sources. By increasing the absorbed energy, heating becomes important, giving rise to induced thermochemical effects (that is, thermal denaturations) that are often useful in biological applications. When the energy density on the target becomes very high, the radiation pressure becomes nonnegligible. Moreover, some nonlinear phenomena may occur, such as optical saturation, two-photon absorption, or dielectric breakdown (at 10^{10} W/cm^2), which lead to ionization and nonselective destruction of the material. *See* CHLOROPHYLL; CYTOCHROME; HEMOGLOBIN; NUCLEIC ACID; PROTEIN.

Laser spectroscopy. This method is used to probe biological processes involving very fast reactions or to study structural changes of complex molecules. The techniques are flash photolysis (in the nanosecond and picosecond range) and Raman spectroscopy.

Many fast reactions can be studied with flash photolysis because lasers provide extremely intense light pulses of very short duration. A large number of excited molecules is therefore produced in a short time. The two main examples of processes studied with this technique are photosynthesis and vision, but the technique has also been applied to study the excited states of many biological molecules. *See* PHOTOCHEMISTRY; PHOTOLYSIS.

Resonance Raman spectroscopy provides information about the vibrational levels of molecules. With lasers used instead of mercury lamps, the quality of the Raman spectrum is improved. *See* RAMAN EFFECT; SPECTROSCOPY; ULTRAFAST MOLECULAR PROCESSES.

Microirradiation of cells. Microirradiation is a useful technique for the study of the photosensitivity of various parts of a single cell. Alteration of a specific organelle after irradiation permits the study of cell functions. This technique has had a wide application in plant cells as well as animal cells. In some cases, animal cells can be stained with supravital dyes to confer absorption at the wavelength of irradiation. Because of the very high power density of the focused laser light, extremely low concentrations of dye can be used, and toxic effects can consequently be avoided. The laser beam is focused through the objective of a microscope onto the cell. The diameter of the beam depends on the divergence of the laser and on the focal length of the objective. In practice, it is easy to obtain spots as small as 1 micrometer in diameter. Ruby, neodymium, and argon lasers are used for this purpose. Ultraviolet excimer lasers are also valuable tools when ultraviolet light and high power density are required.

The organelles which have been the most studied are mitochondria and chromosomes, although other organelles such as mitotic organelles and myofilaments have been irradiated.

Mitochondria. Mitochondria of cells in tissue culture have been irradiated either after staining with Janus green or by using directly the natural absorption of flavins and cytochromes which are present in large amounts in cell types such as heart muscle cells. By using Q-switched lasers, thermal effects are observed in Janus green–stained mitochondria, while with continuous lasers photochemical effects are observed in unstained mitochondria. The possible role of mitochondria in maintaining the beating rhythm of myocardial cells has been investigated in both conditions. *See* MITOCHONDRIA.

Chromosomes. Irradiation of segments of chromosomes is useful for genetic studies. Deletions of small regions of chromosomes are easily obtained after microirradiation with ultraviolet light or visible light plus acridine orange. By this technique, it is possible to assign genes to specific chromosomes, and to generate "mutants" by damaging small selected areas of deoxyribonucleic acid (DNA). Selected parts of chromosomes can be dissected very accurately with an excimer laser into pieces smaller than 0.5 μm in width. Such a technique can be used for microcloning and preparing DNA probes. In addition, location of chromosomes in the interphase nucleus has been performed after ultraviolet or photosensitized damage of chromatin. *See* CHROMOSOME ABERRATION.

Insertion of foreign genetic material. Because self-healing holes in cell walls and membranes can be produced by laser microbeams, micro-irradiation has been applied successfully for the insertion of foreign DNA. In plant cells where conventional techniques can be used only on protoplasts, laser microbeams enable the introduction of DNA without removing the cell wall. The experiments described for mammalian and plant cells clearly indicate that the introduced DNA can be stably incorporated into the genome of the recipient cells.

Optical traps. Optical trapping of biological objects using a laser microbeam allows the manipulation of cells and cellular organelles. This process has been termed optical tweezers. Indeed, significant forces (radiation pressure) are exerted on microscopic objects illuminated by focused beams. By using a 100× objective, cells can be moved over distances greater than 1 mm in the transverse plane, and over about 70 μm in the axial direction.

In addition, optical traps have been used to study about chromosome motion and cell mobility or to manipulate organelles within cells. They have also been applied to selective cell fusion in combination with laser microsurgery.

Confocal microscopy. Confocal fluorescence microscopy allows cellular three-dimensional imaging and analysis of any fluorophore within a cell. Selective fluorophores of cellular components are excited with laser light, and a complete imaging of the cell structure is built up by scanning all points in the field of view. Quantitative analysis of image parameters is achieved by an image processing computer. Therefore, it is possible to use confocal scanning microscopy to record images for analysis in cell biology. *See* CONFOCAL MICROSCOPY.

Flow cytometry. Mammalian cells, treated with appropriate fluorescent stains, are made to flow in single file through a continuous-wave laser light, such as an argon laser which causes them to fluoresce. By measuring the fluorescence emission, it is possible to analyze heterogeneous cell populations, measure physiological cell response, and isolate cell preparations having different properties.

Surgery. Continuous-wave lasers have been employed as a "light knife," that is, as a surgical cutting and coagulation tool. Generally, high-power CO_2 and argon lasers are used for this purpose. The laser beam is delivered through a mirror system located in an articulated arm (CO_2 laser) or through optical fibers (argon laser) to an endoscope, and can be easily directed into various body cavities. When the laser energy is focused onto a tissue surface, a small volume of tissue is heated, and thus only this area is "cut off." An advantage of this procedure is that small capillaries are coagulated, preventing hemorrhage resulting from cut blood vessels and providing better healing. Lasers have been especially used for otorhinolaryngology, gynecology, urology, and abdominal surgery.

Ophthalmology. Argon laser photocoagulators are commonly used for treating retinopathies, particularly for the prevention of the detachment of the retina and for diabetic retinopathies. Treatment of glaucoma and cataract with pulsed lasers has been attempted.

Dermatology. Laser irradiation is used for removal of foreign pigments in the skin (tattoos), for treatment of vascular disorders ("wine marks"), and for removal of various pigmented skin lesions. Because of the higher absorption of laser light in the pigmented spots than in the skin, thermal effects from the absorbed energy cause preferential coagulation of the pigment and the local constriction of blood vessels with little change in the unpigmented skin.

Cancer photochemotherapy. Red laser light has been used to destroy cancer cells which were previously photosensitized with derivatives of hematoporphyrin, a drug which preferentially localizes in malignant cells. The basis of this technique, also called photodynamic therapy, is an oxygen-dependent reaction produced after light excitation of hematoporphyrin derivatives, which leads to the death of the stained cancer cells. Photodynamic therapy is used for the treatment of certain superficial small tumors, especially in the bladder and the bronchi.

A diagnostic possibility to detect cancers is given by the fact that hematoporphyrin derivatives fluoresce when irradiated with blue or ultraviolet light; because the drug concentrates in cancer cells, tissues which fluoresce are probably malignant.

Low-power density effects. Among controversial laser use and effect, acupuncture and so-called biostimulation can be mentioned. Low-power helium-neon laser beams have been used instead of silver or gold needles for acupuncture. In other respects, laser biostimulation at low fluences in the red has been claimed to increase DNA synthesis, collagen synthesis in fibroblasts, and cell proliferation. Consequently, helium-neon lasers have also been employed to accelerate wound healing. *See* LASER.

Giuliana Moreno; Christian Salet

Bibliography. K. Atsumi (ed.), *New Frontiers in Laser Medicine and Surgery*, 1983; R. V. Bensasson, E. J. Land, and T. G. Truscott, *Excited States and Free Radicals in Biology and Medicine*, 1993; M. W. Berns, W. H. Wright, and R. W. Wiegand Steubing, Laser microbeam as a tool in cell biology, *Int. Rev. Cytol.*, 129:1–44, 1991; B. W. Henderson and T. G. Dougherty (eds.), *Photodynamic Therapy*, 1992; G. Weber and K. O. Greulich, Manipulation of cells, organelles and genomes by laser microbeam and optical trap, *Int. Rev. Cytol.*, 133:1–41, 1992; M. L. Wolbarsht (ed.), *Laser Applications in Medicine and Biology*, vol. 4, 2000.

Laser photochemistry

A branch of chemistry in which reactions are induced or altered by laser light. The initial part of any photochemical reaction involves an optical transition to some excited state of molecule. These excited states could involve electronic, vibrational, and rotational excitation. The particular photochemical product that results from the absorption of light depends on the specific excited state species created during the irradiation. Thus, the properties of the light source often determine the photochemical product. Lasers have had an immense impact on the field of photochemistry by providing scientists with an intense, polarized, and nearly monochromatic source of light. There are lasers that extend from wavelengths of less than 110 nanometers (vacuum ultraviolet) to more than 100,000 nm (far infrared); for comparison, the entire visible spectrum extends from only 400 nm (violet) to 700 nm (red).

Advantages of lasers. Use of lasers in photochemical applications provides three main advantages over conventional light sources such as discharge or arc lamps. First, lasers are generally more powerful than conventional light sources. A continuous-wave argon ion laser can produce 10 W at 514.5 nm. Pulsed lasers, which compress the light energy into very short time periods (10^{-6} to 10^{-12} s), can generate correspondingly higher peak powers, typically from 10^3 to 10^9 W. These very short, intense pulses of light are used by the photochemist to monitor the time evolution of excited-state species or the appearance rate of products in a photochemical reaction.

Second, laser light can be collimated into a beam with a very small divergence angle, routinely less than 1/100 of a degree. The high degree of collimation of laser light permits efficient illumination at a chosen point within the sample, which could be far from the light source. This collimation is important in photochemistry. It permits the photochemist to confine the photochemical activation to some very small and precisely located area, such as in the fabrication and repair of microelectronic devices. In addition, the photochemical event, which has a very low absorption cross section, can be induced by photolyzing a very small region of the chemical sample with the total laser output power.

Third, laser light is exceptionally pure in color. The spectral purity of light can be described by a band width measured in wavenumbers (cm^{-1}), defined as the frequency width divided by the speed of light. While the full visible spectrum is 10,000 cm^{-1} wide, a typical laser may be as narrow as a few cm^{-1}. This particular advantage of laser light has had its largest impact in photochemistry. It is difficult to obtain spectral purity while maintaining high power with conventional light sources. Consequently, when conventional sources are used to irradiate molecules that have many possible excited-state transitions, a distribution of excited molecules results. These molecules may have been excited to several different electronic states with many possible vibrational or rotational energies. Under such conditions, it is often impossible to identify the excited electronic states that produced the various photochemical products. The high spectral purity of the laser obviates this problem, and so the identity of the excited electronic state is almost always known to the laser photochemist.

The high spectral purity can be a disadvantage when the laser wavelength does not correspond to the required absorption wavelength for the molecule under study. This difficulty has been largely solved in the visible range, where organic dye lasers permit continuous tuning over a significant portion of the spectrum. Nonlinear, pulsed laser techniques, such as frequency doubling, metal-vapor four-wave mixing, stimulated Raman scattering, and optical parametric oscillation, have extended the accessible frequencies to include the vacuum ultraviolet and near-infrared region of the spectrum. *F*-center, diode lasers, and difference-frequency techniques generate continuous-wave radiation over a large portion of the near infrared.

Because of the advantages of lasers, there is a constant demand for the development of lasers that will serve in new regions of the spectrum. In spite of the advantages of lasers, conventional light sources are still widely used in most photochemical investigations and commercial processes because of their simplicity and low cost. *See* COLOR CENTERS; RAMAN EFFECT.

The two general areas of laser-induced photochemistry are vibrational photochemistry and electronic photochemistry. In vibrational photochemistry, the chemical reaction occurs entirely on the ground electronic state of the molecule, whereas electronic photochemistry occurs via some excited electronic state, usually involving the first excited singlet state (all electrons spin-paired) and possibly the lowest-lying triplet state (two electrons spin-unpaired) for most polyatomic molecules. Electronic photochemistry is subject to the selection rules that govern optical electronic transitions. Although electronic photochemistry has been profoundly affected by the advent of lasers, vibrational photochemistry, which requires intense light sources, is not even possible without lasers. *See* ELECTRON SPIN; TRIPLET STATE.

Vibrational photochemistry. In vibrational photochemistry, the chemical reaction occurs on the ground electronic state of both the reactant and product. With this technique, the laser light excites molecules to energetic vibrational states lying above the reaction barrier. By varying the total vibrational energy within the molecule, the photochemist can learn about the energy dependence of the collisional deactivation process and the rate constant for reaction. Laser photochemistry can be used to investigate the interplay between the thermochemical factors and the rate constants of reactions that can also be initiated by thermal activation. This is particularly interesting when there are competitive reaction channels.

Infrared multiphoton photochemistry. Both isomerization and dissociation reactions can be induced with intense infrared lasers. In this photoactivation technique, which has been used to study many different reactions, the first few photons are absorbed more or less resonantly by vibration-rotation states. It is here that it becomes possible to selectively excite one isotope within an isotopic mixture. The excitation scheme for multiphoton and single-photon methods is shown in **Fig. 1**. *See* PHOTON.

As the molecule becomes more energetic, the number of possible vibration-rotation states increases sharply. This is known as the vibrational quasi-continuum. Here the subsequent absorption cross sections are small; but with an intense carbon dioxide (CO_2) laser the molecule will continue to absorb photons and eventually become energetic enough to react. It is often difficult to determine the exact number of photons absorbed by the molecules. Indeed, sometimes the molecules are overexcited beyond the activation barrier so that the reaction rate exceeds the thermal rate. If two dissociation

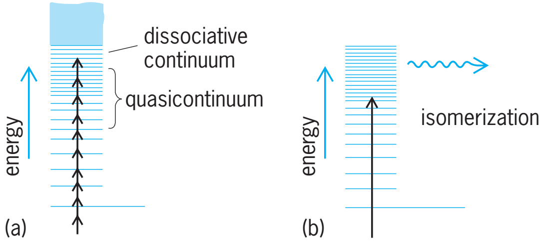


Fig. 1. Excitation scheme using intense infrared lasers. (*a*) Multiphoton. (*b*) Single photon. The black arrows represent the absorption of a photon.

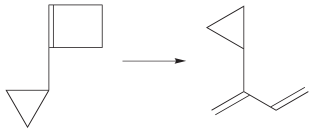
reactions have similar activation barriers, it is possible to change the ratio of the two products by increasing the excitation rate of the laser. For example, there are two dissociation channels for difluorodichloromethane (CF_2Cl_2) that lie close in energy, as shown in reaction (1),

$$CF_2Cl_2 \begin{array}{l} \nearrow CF_2Cl + Cl \quad \Delta H = 80 \text{ kcal/mole} \\ \searrow CF_2 + Cl_2 \quad \Delta H = 71 \text{ kcal/mole} \end{array} \qquad (1)$$

where ΔH = enthalpy change. When this reaction is initiated by laser multiphoton infrared excitation, the relative yield of $CF_2\cdot$, as measured by resonant fluorescence, depends on the total reaction yield. As the total reaction yield is increased, the amount of $CF_2\cdot$ formed by multiphoton dissociation first increases, then decreases. This change in the relative product yield indicates that the reactant was overexcited above the 71 kcal/mole barrier to a level of excitation where both reaction channels can compete. *See* ENTHALPY.

Single-photon photochemistry (overtone excitation). A method of photoactivation known as laser vibrational overtone activation induces highly excited vibrational states within a molecule by single-photon absorption. In contrast to multiphoton excitation, the exact energy of the activated molecule is known from the wavelength of the excitation laser. Because the molar extinction coefficient ϵ is very small ($\epsilon \simeq 10^{-4}$ or 10^{-5} in contrast to normal allowed electronic transitions, where $\epsilon \simeq 10^4$ to 10^6, or typical infrared transitions, where $\epsilon \simeq 1$ or 10), the gaseous sample is placed within the laser cavity, where the number of photons is a thousand times higher than the laser output. Typically, a tunable dye laser operating in the visible region is tuned to the vibrational overtone transition wavelength, and the photolysis products are detected by gas chromatography after a timed photolysis period. Since each type of carbon-to-hydrogen (C-H) bond (methyl, vinyl, methylenic, and acetylenic) in the vibrational overtone spectra of hydrocarbons absorbs at a characteristic frequency, the specific moieties within the molecule can be excited by tuning the laser wavelength. *See* GAS CHROMATOGRAPHY; PHOTOLYSIS.

Photochemists hoped that laser vibrational overtone activation would provide the long-sought mode-selective chemistry. The idea of mode-selective chemistry involves forcing the molecule to undergo one reaction rather than another by tuning the laser wavelength from one vibrational absorption feature to another. Many unsuccessful attempts have been made to achieve this. One elegant experiment involved 1-cyclopropylcyclobutene (**1**), which isomerizes to 2-cyclopropyl-1,3-butadiene (**2**) as shown in reaction (2).



(2)

(**1**) (**2**)

When cyclobutene is excited into the sixth vibrational level of the methylenic C-H bond, the rate constant is 10^8 s^{-1}, but the reaction rate constant for 1-cyclopropylcyclobutene is 10^5 s^{-1}. No difference in rate was observed for this reaction when C-H bonds on the cyclopropyl ring or the cyclobutene ring were excited, indicating that mode selective chemistry was not achieved.

Deanne Snavely

Mode-selective photochemistry. Recently, scientists have become interested in trying to use lasers to cause selective photochemistry. In particular, they would like to break a specific bond in a molecule while leaving the other, weaker bonds unbroken, or to cause a reaction at a specific site in a molecule. This area of research is called mode-selective photochemistry.

Several schemes for mode-selective photochemistry have been tried, and two have shown promise. The first is direct excitation of a molecule with ultraviolet photons to produce an electronic state that correlates to a particular set of reaction products. The second is to vibrationally excite a molecule using visible or infrared photons to energize a particular bond in a small molecule. If the energy can be localized in the bond, it will take less additional energy to break it, and the bond can be thought of as having been weakened. Mode-selective bimolecular photochemistry has also been demonstrated with this second technique.

Laurie Butler, University of Chicago, has used ultraviolet radiation to demonstrate mode-selective photochemistry by using molecules that contain two different halogen atoms bonded to a carbon atom, that is, two chemically similar bonds of differing strengths. Butler has been able to break the stronger C-Br bond of a molecule in the presence of a weaker C-I bond with nearly unit efficiency. **Figure 2** is a schematic of an experiment performed using

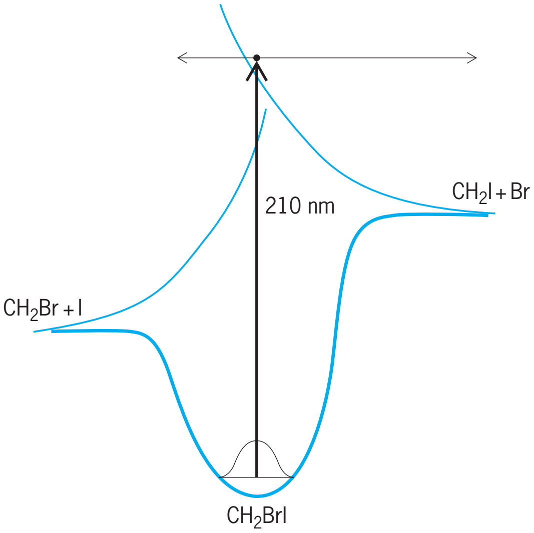


Fig. 2. Schematic representation of a photodissociation experiment where 210-nm light is used to excite bromoiodomethane, causing rupture of the strong C-Br bond in preference to the weaker C-I bond.

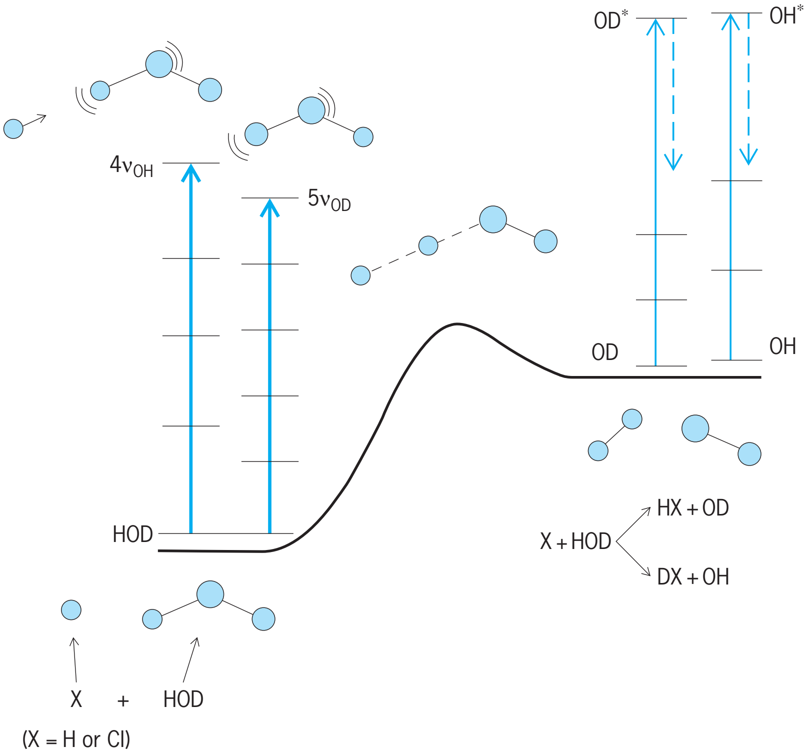


Fig. 3. Schematic drawing of the reaction barrier for breaking either the O-H or the O-D bond in the reaction of H + HOD or Cl + HOD. The thick arrows represent vibrational overtone excitation of either the $4\nu_{OH}$ or the $5\nu_{OD}$ bonds of HOD. The thin arrows represent the use of ultraviolet laser light to detect the OD or OH products.

ultraviolet light near 210 nanometers to excite bromoiodomethane. This wavelength of light will excite an electron to an antibonding orbital that correlates to products CH_2I + Br. At a lower energy, electronic states would couple to the lower-energy-channel products CH_2Br + I. Because the dissociation is very rapid, the energy does not have time to redistribute in the molecule, and only CH_2I + Br products are observed. *See* HALOGEN ELEMENTS.

The idea of using vibrational excitation to selectively weaken a bond has been tried for many years. After the invention, in the 1970s, of intense carbon dioxide lasers, which produced infrared radiation near 10 micrometers, many groups attempted to break particular bonds in molecules by resonantly exciting low-frequency vibrations. This approach was not successful because the energy would rapidly flow throughout the molecule, a process that physical chemists call intramolecular vibrational redistribution, and the weakest bond would predominantly break. With the advent of powerful visible lasers, in the 1980s, scientists began trying to use a single energetic photon to induce vibrational mode-selective photochemistry. They would tune the laser to be resonant with a weak vibrational overtone transition (a transition where the vibrational quantum number changes by more than one) in a molecule possessing a low-energy reaction process, such as isomerization or dissociation. The hope was that the single photon would deposit energy in the molecule with enough specificity to accelerate a particular reaction so that selective chemistry could be observed. Once again, it was found that molecules large enough to have low-lying reaction barriers had a very high density of vibrational states and strong couplings between them, so that the photon energy experienced intramolecular vibrational redistribution, and for the most part statistical behavior was observed.

Fleming Crim, University of Wisconsin, was able to succeed with this approach in small molecules that have a low density of states and therefore little intramolecular vibrational redistribution. In particular, Crim was able to demonstrate a high degree of mode-selective photodissociation using a technique he calls vibrationally mediated photodissociation. In the first step of this process, a small molecule, such as water (HOD) or hydrogen peroxide (HOOH), is excited via an overtone transition. This vibrational excitation does not possess enough energy to cause dissociation, but the further excitation of the molecule from the vibrationally excited state to a dissociative electronic state can be very efficiently and selectively achieved. In this manner, only the initially excited molecules are dissociated, and they dissociate along the initially excited bond. For example, exciting either the OH stretching vibration or the corresponding OD vibration will cleave either the H or the D of the HOD, and in the case of hydrogen peroxide (HOOH) will cause the O-H bond to break despite the presence of the weaker O-O bond.

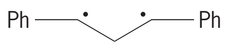
Bimolecular mode-selective photochemistry has also been demonstrated by Crim (**Fig. 3**). Once again, an overtone vibrational excitation of one of the bonds of HOD is used to energize a bond. An H atom or Cl atom is produced concurrent with the vibrational excitation of the HOD molecules. A collision between a vibrationally excited HOD molecule and an H atom or a Cl atom has enough energy to surmount the barrier to this endothermic reaction. The exciting aspect of this experiment is that the selectivity of the initial vibrational excitation is not lost in the collision that produces the reaction. If the O-H bond is initially excited, then H_2 + OD or HCl + OD are the final products; and if the O-D bond is initially excited, HD + OH or DCl + OH are the final products.

These experiments show that mode-selective photochemistry is not only possible but can be made to be extremely efficient.

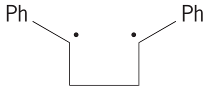
David W. Chandler

Time-resolved photochemistry. An important advantage of lasers in photochemistry is their ability to time-resolve the chemical processes that occur after the photon is absorbed. Transient absorption and time-resolved laser-induced emission can be used to identify the species involved in the reaction and to determine lifetimes. For applications in organic chemistry, a typical laser flash photolysis experiment employs a nanosecond laser pulse from a nitrogen or excimer laser (perhaps pumping a dye laser) to initiate the chemical reaction and a tunable light source (perhaps a pulsed xenon lamp on another laser) to monitor the transient species. The time resolution of the experiment is determined by the duration of the light source and the response time of the detector. The overall time resolution for this technique is usually 3–6 nanoseconds.

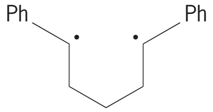
Laser flash photolysis has been used to measure the lifetimes of many biradical species. In many common organic solvents, these lifetimes range from a few nanoseconds to several hundred nanoseconds. For instance, the biradical lifetimes of structures (3)–(6)



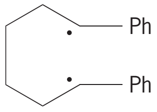
Ph = phenyl;
τ = 20
(3)



τ = 190
(4)



τ = 920
(5)



τ = 1080
(6)

as measured by laser flash photolysis in acetonitrile increases from 20 to 1080 ns. The lifetimes increase as the distance between the radical centers increases. Intersystem crossing occurs when an excited molecule in a singlet state converts to an electronic triplet state. Radicals (3)–(6) must undergo intersystem crossing before they undergo further chemical reaction. However, for these species this intersystem crossing rate is sufficiently slow to be the rate-determining step. As the distance between the radical centers increases, the rate of intersystem crossing increases, which in turn lengthens the lifetime of the transient diradical. *See* FREE RADICAL; LASER SPECTROSCOPY.

Microelectronics fabrication. There are many advantages of using laser photochemistry to fabricate microelectronic devices. Many of the materials (mercury-cadmium-tellurium, gallium arsenide, and plastics) used in microelectronics are sensitive to thermal decomposition; consequently, it is desirable to maintain lower temperatures during fabrication. With laser processing, no bulk heating occurs, since the chemical activation energy is provided by the laser photon. In addition, the laser beam can be focused to process a very small area on the substrate surface, so that only localized heating is created. The dimensions of the metal lines deposited on the substrate surface must be very narrow (on the order of 1 micrometer); the lower temperatures avoid metal diffusion across the surface, broadening the deposited line. An important advantage of lasers is the generation of patterns of deposited material without the need for masking. In conventional microelectronics fabrication, a patterned mask is usually deposited onto the substrate surface (typically silicon). This mask is then used to transfer the pattern onto the underlying surface, which might be further processed by etching in subsequent steps. With laser deposition and laser-driven doping, the same result can be achieved in a few steps.

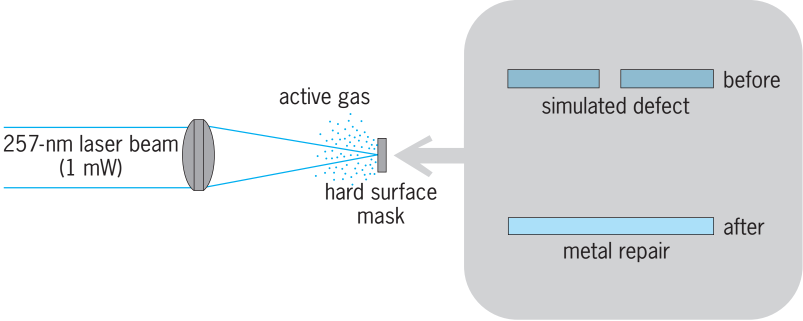


Fig. 4. Mask repair by laser deposition using a low-power laser at 1-μm resolution. (*After K. G. Ibbs and R. M. Osgood, eds., Laser Chemical Processing for Microelectronics, Cambridge University Press, 1989*)

Laser chemical deposition of metal is used in mask repair. This process involves repairing a discrete defect in a metal line on a semiconductor surface (**Fig. 4**). For example, metallic cadmium can be deposited by photodissociating gaseous dimethylcadmium with a weakly focused, frequency-doubled argon ion laser operating at 257 nm. Laser focus can be controlled to cover a 10-μm-diameter area, small enough to patch the defective cadmium line precisely.

The interference pattern generated by two laser beams at a semiconductor surface can generate a diffraction grating by using a technique known as laser-induced wet etching. In this process the semiconductor is typically submerged in an acidic aqueous solution. Light-induced electrochemical reactions involving the metal atoms of the semiconductor render the illuminated portion more soluble in the aqueous solution. Consequently, etching of the surface occurs only in the illuminated area and produces the interference pattern formed by the laser beams. *See* INTEGRATED CIRCUITS; VAPOR DEPOSITION.

Femtosecond photochemistry. Chemical dynamics is an area of chemistry that attempts to determine precisely how an individual molecule undergoes a chemical reaction. This involves preparing the molecular sample in some initial quantum-mechanical state (or some narrow distribution of initial states) and determining the final states that are produced after a chemical reaction. The most widely used technique for producing narrow initial-state populations is the supersonic expansion. A supersonic jet expansion is produced by expanding a mixture of the photochemically active species and a diluent gas such as nitrogen (N_2) or helium (He) from high pressure to low pressure through a small orifice. As the gaseous mixture expands adiabatically into the low-pressure region, it cools substantially. If helium is used, temperatures of a few kelvins are obtained routinely (**Fig. 5**).

Laser photoexcitation techniques such as wavelength selection or polarization alignment further limit the number of initial states for the given photochemical reaction. The initial relative translational energy of reagents in a bimolecular reaction can be controlled also by using two supersonic jets intersecting at a particular angle. Pulsed lasers having

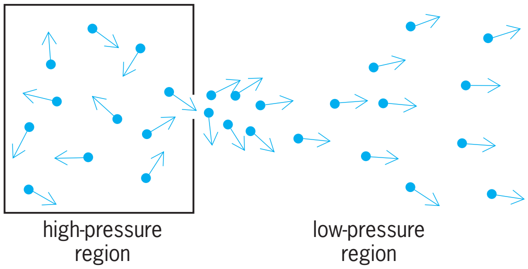


Fig. 5. Diagram of the expansion of a gas in a supersonic jet. The gas cools as it expands into the low-pressure regions.

nanosecond and picosecond pulse duration are used in chemical dynamics to keep the time intervals between the initiation of the reaction and the diagnostics of the final states short to avoid contamination of subsequent chemical reactions or changes of final state by collisions.

As the ability to control the time pulsewidth of lasers has moved into the femtosecond (10^{-15}) range, another area of chemical dynamics has opened—that of observing a chemical reaction while it is occurring. For example, it takes about 200 fs for the dissociation reaction of the triatomic molecule cyanogen iodide (ICN) to occur. If this reaction is initiated by a laser with 6-fs pulses and probed at short intervals after the dissociating pulse, the progress of the reaction through the transition state can be monitored. The time interval of 100 fs between the laser pulse initiating the dissociation reaction and the pulse monitoring the progress of the reaction is generated by sending the monitoring pulse through an extra distance of 3×10^{-5} m on its path to the ICN sample. (Since light travels 3×10^{-7} m/fs, this extra distance introduces a 100-fs delay.) In reaction scheme (3) the excitation laser at 307 nm excites the

$$\mathrm{ICN}^* \longrightarrow [\mathrm{I}\cdot\cdot\cdot\cdot\cdot\mathrm{CN}]^* \longrightarrow \mathrm{I} + \mathrm{CN} \qquad (3)$$

ICN to ICN* on the repulsive excited-state surface labeled V_1 in **Fig. 6**. The delayed probe laser, operating at 388.5 nm, excites the free CN fragment, from which fluorescence is detected. The delay between the excitation laser at 307 nm and the probe laser at 388.5 nm measures the time for the free CN fragment to be generated. Furthermore, the progress of the reaction can be monitored by tuning the probe laser wavelength to longer wavelengths. As can be seen in Fig. 4, the energy difference between V_1 and V_2 depends on the separation of I and CN. This means that if the probe laser is shifted to the red of the free CN absorption wavelength, the CN fragment can be excited as it is being generated. This is the first time that observations have been made of a primary chemical reaction in progress and the time required for the fragments to move along their reaction trajectory has been determined. These techniques are under development for application to more complicated reactions. *See* CHEMICAL DYNAMICS; FLUORESCENCE; LASER; PHOTOCHEMISTRY; ULTRAFAST MOLECULAR PROCESSES.

Deanne Snavely

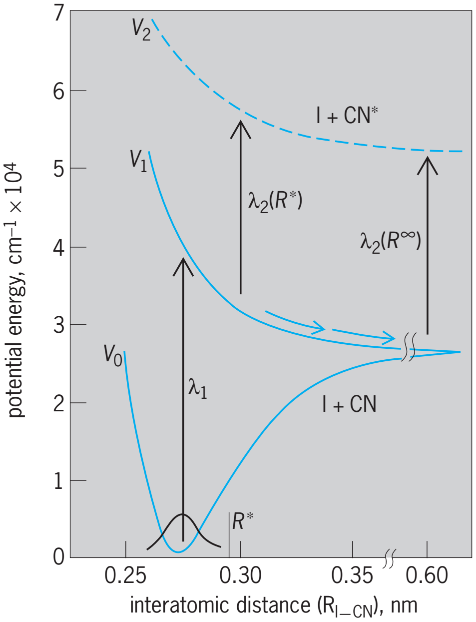


Fig. 6. Potential energy diagram for the dissociation of ICN. (*After A. H. Zewail, Chem. Eng. News, 66(42):24–42, 1988*)

Bibliography. V. N. Bagratashvili et al., *Multiple Photon Infrared Laser Photophysics and Photochemistry*, 1985; L. J. Butler, Chemical reactions beyond the Born-Oppenheimer approximation, *Annu. Rev. Phys. Chem.*, 49:125–171, 1998; J. Chaiken (ed.), *Laser Chemistry of Organometallics*, 1993; F. F. Crim, Bond-selective chemistry: Vibrational state control of photodissociation and bimolecular reaction, *J. Phys. Chem.*, 100:12725–12734, 1996; K. G. Ibbs and R. M. Osgood (eds.), *Laser Chemical Processing for Microelectronics*, 1989; V. Ramamurthy, *Organic Photochemistry*, 1997; H.-G. Rubahn, *Laser Application in Surface Science and Technology*, 1999; C. E. Wayne and R. P. Wayne, *Photochemistry*, 1996.

Laser-solid interactions

Interactions of laser light with solids. The term usually refers to the thermal effects of absorption of high-intensity laser beams. For nonthermal laser interactions with matter *see* LASER PHOTOCHEMISTRY; LASER SPECTROSCOPY; NONLINEAR OPTICS.

One of the unique features of lasers is their ability to concentrate the energy of light so that a surface can be exposed to a photon flux of 10^6 to 10^{12} W/cm^2. These extraordinarily high power densities are attainable even with lasers of moderate size and average output of only a few watts, since light can be emitted in pulses of a few nanoseconds or even less than 1 picosecond duration and a coherent beam can be focused into spots as small as 1 micrometer in diameter. The high power densities allow melting and even vaporization of any solid material that is sufficiently opaque at a given wavelength or photon energy. This has led to a number of applications involving cutting and drilling of ceramics and other brittle materials, even diamonds. Welding of components from the smallest wires to huge steel plates is done commercially with high-power lasers. Metal alloying in surface regions is also a domain of lasers. Modification of semiconductors with intense laser irradiations has emerged as one of the most exciting fields of scientific discovery while at the same time promising substantial practical benefits for the semiconductor device industry. *See* LASER ALLOYING; LASER WELDING; SEMICONDUCTOR.

Photon-electron-phonon interactions. Photons with energy exceeding the fundamental band-gap energy E_g of the semiconductor are absorbed near the surface, exciting electrons from the valence to the conduction band and forming electron-hole pairs. For direct band-gap semiconductors, such as gallium arsenide (GaAs), indium phosphide (InP), and indium antimonide (InSb), light is absorbed within 100 nanometers of the surface. For indirect band-gap semiconductors, such as silicon and germanium, the transition from transparency to absorption is a more gradual function of the light wavelength, because for each transfer of energy from a photon to an electron, a phonon has to be created or annihilated to assure conservation of momentum. *See* BAND THEORY OF SOLIDS; CRYSTAL ABSORPTION SPECTRA.

For photon energies less than E_g, semiconductors are largely transparent, at least at moderate light intensities. Some photons are still absorbed by exciting electrons and holes within the conduction and valence bands to higher energies (free-carrier absorption), but unless the semiconductor is heavily doped with impurities that lead to a large concentration of free carriers, the light beam can penetrate to a depth of 0.04 in. (1 mm) or more before it is fully absorbed. Far-infrared light can interact directly with the lattice vibrations or phonons, that is, the energy of photons can be directly transformed into heat, but this absorption mechanism is quite weak. The low absorption channels are often very important at high photon fluxes, as they can initiate surface heating that tends to make solids more opaque. This feedback mechanism facilitates laser melting and vaporization of almost any solid substance. In semiconductors, free-carrier absorption by electrons and holes generated by optically and thermally stimulated band-to-band transitions contributes a large component to the total absorption at high power densities. *See* LATTICE VIBRATIONS; PHONON.

Excited electrons and holes share energy with one another very rapidly through electron-electron and electron-plasmon interactions. They release energy to other electric carriers and to phonons until they reach the lowest energy possible within their respective energy bands. Then they recombine by one of several mechanisms: radiatively by emitting a photon, nonradiatively by transferring the amount of energy E_g into heat or by an Auger process that transfers the energy to a third particle (electron or hole), exciting it to an energy higher by E_g (**Fig. 1**). Auger interactions are very important for a high density of photoexcited carriers. Although they do not reduce the energy of the electronic system, they convert the band-gap energy into kinetic energy of carriers that can be dissipated rapidly into heat by phonon emission. All the photon energy absorbed by the solid is transformed into heat in about 1 ps, that is, almost instantaneously. *See* AUGER EFFECT.

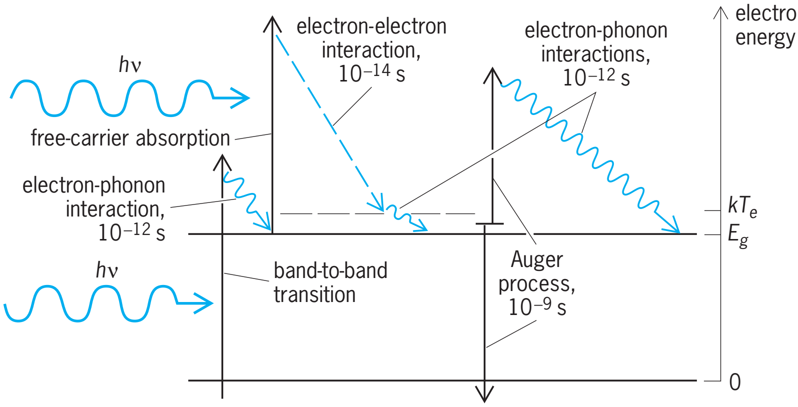


Fig. 1. Absorption of light and energy transfer mechanism for a very intense flux (10^7-10^9 W/cm^2) of photons with energy $h\nu$ exceeding the band-gap energy E_g. Approximate times for processes to take place are given. kT_e is the average thermal energy per electron, the product of Boltzmann's constant k and the electron temperature T_e.

Annealing of ion-implanted layers. Ion implantation has become a dominant method of introducing controlled quantities of impurities near the surface of silicon and other semiconductors for the purpose of adjusting their electrical conductivity and forming diodes, transistors, and entire integrated circuits. The implanted layers need a heat treatment to repair the displacement damage caused by bombardment with energetic ions and to move the implanted impurity ions into lattice locations where they replace host atoms and become electrically active. Laser heating is particularly suitable for annealing since only the implanted regions are heated. This capability of delivering heat selectively to a very thin layer is the basis of "laser annealing" of semiconductors. *See* ION IMPLANTATION.

Pulsed laser annealing. A short, high-intensity laser pulse melts the layer of silicon damaged by ion implantation. As soon as the pulse is over, the layer freezes, starting at the unmelted single crystal (**Fig. 2**). The single crystal forms a template on which the melt solidifies, preserving the perfection of the template. This process of liquid-phase epitaxy is the same as conventional crystal growth from the melt, except that the solid-liquid interface moves at a velocity of many feet per second, driven by the temperature gradients of about 10^7 degrees per inch. The quality of the crystalline structure is not diminished by the rate of growth, but some new phenomena appear. For example, most impurities that are present in the liquid are incorporated into the solid more completely than during slow crystal growth. Maximum concentrations of some impurities in the crystalline semiconductor can exceed by orders of magnitude the near-equilibrium solid solubilities.

Pulse irradiation allows investigation of the physical limits of crystal growth and the thermodynamic properties of different phases. For example, it was

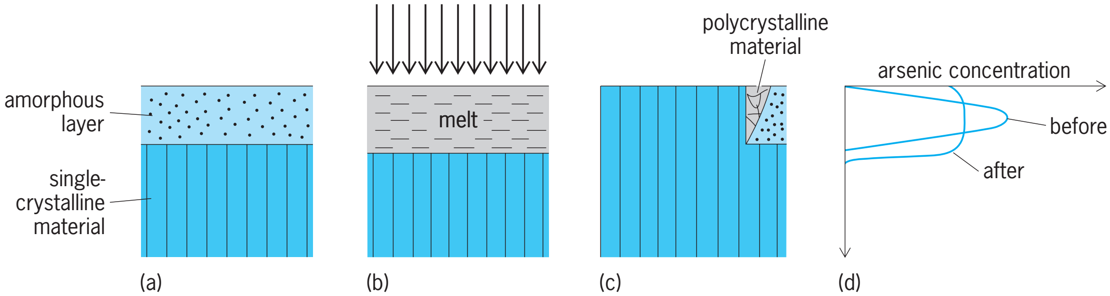


Fig. 2. Pulsed laser annealing by liquid-phase epitaxy. (*a*) Cross section of a crystalline semiconductor with a thin (typically 100 nm thick) amorphous surface layer formed by ion implantation. (*b*) The entire thickness of the amorphous film and some of the underlying crystal melted by a laser pulse. (*c*) Molten layer recrystallized from the crystalline template to the surface. At the edge of the laser-irradiated spot, where the melt has not reached the crystalline substrate, the material is polycrystalline. (*d*) Typical depth profiles of arsenic implanted into silicon, before and after pulsed laser annealing.

determined that amorphous silicon melts at a lower temperature than crystalline silicon, confirming theoretical predictions based on Gibbs free-energy considerations. When the solidification rate of molten semiconductors exceeds about 50 ft/s (15 m/s; this is accomplished with pulses shorter than 2 ns and short-wavelength irradiation), the liquid is supercooled so much that it transforms into an amorphous layer instead of a single crystalline layer. Therefore, depending on irradiation conditions, amorphous → crystalline and crystalline → amorphous phase transitions are achievable, even with the same laser.

Continuous-wave. Laser beams can anneal implantation displacement damage without melting, by solid-phase epitaxy. At elevated temperature the atoms rearrange themselves into an orderly crystalline structure. The process starts at the boundary between crystalline and damaged regions, and this interface moves through the amorphous film (**Fig. 3**), at a velocity determined by the activation energy E_a (for silicon $E_a = 2.7$ eV). Typically about 1 millisecond time is required to regrow ion-implanted surface layers. Pulses of such duration could be used, but better results are obtained with a continuous-wave laser beam scanned across the surface at a velocity that yields about 1 ms dwell time. This type of laser annealing is physically equivalent to furnace annealing, but there are some important differences. Heating is so brief that practically no diffusion of impurities takes place. The impurity profiles after annealing are indistinguishable from as-implanted depth profiles. In contrast, furnace annealing at high temperatures that are necessary for complete defect removal causes significant diffusion of doping impurities. Laser heating is localized both in depth and in a lateral direction. Small areas can be selectively annealed with a tightly focused beam, and micrometer-size patterns can be written in implantation amorphized material. Since high surface temperatures are attained with a laser in less than 1 ms, measurements of solid-phase epitaxial regrowth rates up to the melting temperature became possible, broadening the understanding of solid-phase epitaxy. *See* CRYSTAL GROWTH.

Silicon on insulator. Thin films of single-crystalline silicon over an insulating substrate are very attractive for high-speed integrated circuits. An important approach to the formation of such films is the controlled melting of thin polycrystalline layers deposited over fused silica substrates or over oxidized silicon wafers. Through a careful control of temperature gradients around the molten spot, by shaping the laser beam or patterning the film, single-crystalline regions can be obtained. One particularly successful approach involves growing single-crystalline films from openings in the oxide where the melt is in contact with the single-crystalline template (**Fig. 4**). The formation of silicon-on-insulator structures will lead in the future to three-dimensional circuits, with several levels of transistors on the same chip. *See* INTEGRATED CIRCUITS.

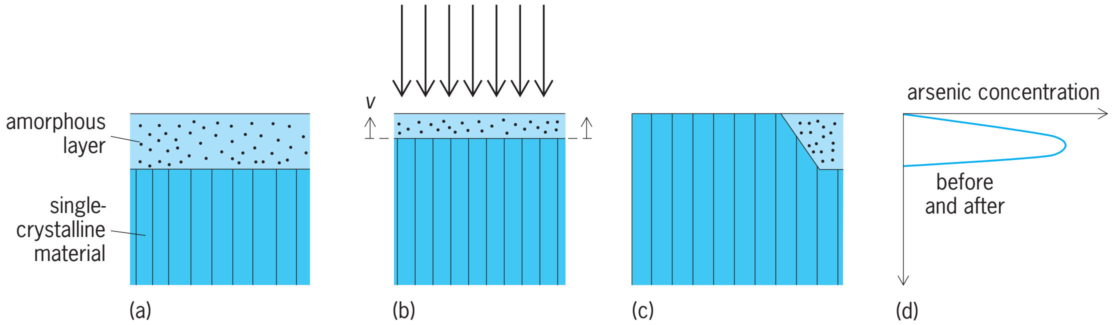


Fig. 3. Continuous-wave laser annealing by solid-phase epitaxy. (*a*) Cross section of the starting structure. (*b*) Crystalline-amorphous interface propagating toward the surface with a velocity *v* that depends on temperature induced by the laser beam. (*c*) The irradiated region fully recrystallized, except at the edge of the laser-heated spot where the interface has a wedge shape. (*d*) Depth profiles of arsenic implanted into silicon, the same before and after continuous-wave laser annealing.

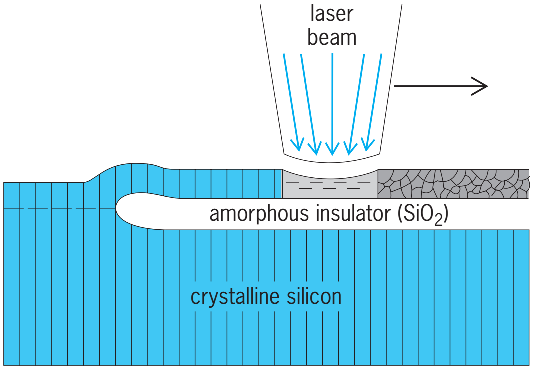


Fig. 4. Melting with a scanning continuous-wave laser beam, allowing lateral epitaxial regrowth of crystalline silicon over a layer of silicon dioxide (SiO$_2$).

Laser-induced-damage gettering. At high power densities, Q-switched laser pulses can vaporize material from the surface of a solid and form microscopic craters. Shock waves associated with vaporization stress the surface and lead to formation of dislocations. The laser-induced surface damage helps in removing undesirable impurities from the device regions of silicon to an inactive area such as the back side of a wafer. The fast-diffusing impurities become trapped or gettered at the damage sites, from where they do not affect device operation.

Laser deposition. Vaporization of materials with nanosecond laser pulses can be used to grow thin films. Each laser pulse removes a small amount of material from the irradiated surface, with the plume of atoms, ions, and small atomic clusters spreading out and condensing on the surrounding surfaces. To obtain controlled deposition, the target and the object to be coated are placed in a vacuum chamber, with the laser beam entering through a transparent window. Films of amorphous silicon and of refractory materials have been formed in this way. Laser deposition has been successfully applied to formation of superconducting films of $YBa_2Cu_3O_x$ with a critical temperature T_c of 90 K. (This means that the film is superconducting up to 90 K or −298°F.)

Laser generation of x-rays. Very high-intensity, short laser pulses focused on a metallic alloy generate temperatures of many million degrees, vaporizing the surface and forming a plume of highly ionized gas. The ions relax to a lower energy by emitting x-ray radiation. These short bursts of x-rays, with typical wavelengths of 0.6 to 2 nm, find application in x-ray lithography, a process that is necessary for making complex semiconductor chips with millions of transistors, each one smaller than 0.5 micrometer.

Inertial confinement fusion. Gigantic lasers, with multiple beams, are used to implode micrometer-size pellets of deuterium and tritium, producing temperatures and pressures equivalent to those inside the stars. The goal is to overcome repulsion of nuclei and fuse them into helium atoms, releasing controllable amounts of thermonuclear fusion energy. *See* LASER; NUCLEAR FUSION. George K. Celler

Bibliography. G. K. Celler, *Modification of Silicon Properties with Lasers, Electron Beams, and Incoherent Light*, CRC Critical Reviews in Solid State and Material Science, vol. 12, pp. 193–265, 1984; S. S. Charschan, *Lasers in Industry*, 1972; *Materials Research Society Proceedings*, annual volumes since 1979; J. M. Poate and J. W. Mayer, *Laser Annealing of Semiconductors*, 1982.

Laser spectroscopy

Spectroscopy with laser light or, more generally, studies of the interaction between laser radiation and matter. Lasers have led to a revolution of classical spectroscopy, because laser light can far surpass the light from other sources in brightness, spectral purity, and directionality; and if required, laser light can be produced in extremely intense and short pulses, far enhancing the possibility of observing nonlinear phenomena. The use of lasers can greatly increase the resolution and sensitivity of conventional spectroscopic techniques, such as absorption spectroscopy, fluorescence spectroscopy, or Raman spectroscopy. Moreover, interesting phenomena have become observable in the resonant interaction of intense coherent laser light with matter. Some of these effects have become the basis for powerful spectroscopic methods, which offer unprecedented spectral resolution, or which permit the investigation of properties of matter that could not be observed previously. Laser spectroscopy has become a wide and diverse field, with applications in numerous areas of physics, chemistry, and biology.

Tunable sources. Early lasers, such as ruby lasers or helium-neon lasers, worked only at a few discrete wavelengths determined by narrow spectral lines of the active medium. The vigorous advances of laser spectroscopy since about 1970 have been largely due to the development of laser sources which are highly monochromatic and in which the wavelength can be tuned continuously over a wide spectral range.

Tunable coherent sources are available for any wavelength region from the far infrared into the near-vacuum ultraviolet. Tunable infrared lasers include high-pressure molecular gas lasers, semiconductor diode lasers, spin-flip Raman lasers, and color-center lasers. In the visible region, organic dye lasers and later titanium:sapphire lasers have proved particularly versatile and powerful spectroscopic sources. Both pulsed and continuous-wave dye lasers cover the entire visible spectrum, including the bordering near-infrared and near-ultraviolet regions. Titanium:sapphire lasers allow a huge tunable range in the red and near infrared when used as continuous-wave sources. Mode-locked titanium:sapphire systems have allowed the realization of ultrashort optical pulses well down in the femtosecond region. Intense shorter-wave ultraviolet radiation can be generated with excimer lasers over limited regions. Semiconductor diode lasers now cover almost all the wavelengths from ultraviolet to far infrared, allowing great reduction in costs and size of experiments.

The generation of harmonic frequencies or sum frequencies in nonlinear optical crystals or gases provides a valuable method to produce tunable, coherent ultraviolet radiation from laser light of longer wavelengths. Optical parametric oscillators and difference-frequency crystal mixers offer corresponding alternatives for the generation of coherent infrared radiation. Stimulated Raman scattering can be used to shift the frequency of a tunable laser by some integer multiple of a molecular vibration frequency and thus to extend the tuning range. These nonlinear frequency-mixing and shifting techniques work best with the intense radiation from pulsed lasers. *See* NONLINEAR OPTICS.

New techniques have been developed to take advantage of laser coherence and spectral purity. Considerable engineering efforts have been devoted to the active frequency stabilization of continuous-wave lasers. By electronically locking the frequency of a continuous-wave titanium:sapphire laser to some reference interferometer, for instance, a linewidth of less than 1 MHz is achieved in commercial systems, and linewidths below 1 Hz have been produced with more sophisticated feedback controls or a resolution on the order of 1 part in 10^{15}. Femtosecond pulsed lasers have been used to create a link between microwave frequencies and optical frequencies in order to take full advantage of the narrow linewidth of tunable lasers.

Absorption spectroscopy. Lasers can replace conventional light sources and spectrographs or monochromators for absorption spectroscopy. The spectral purity of laser light can eliminate instrumental resolution limits, the high intensity helps to overcome detector noise, and the good directionality permits long or folded absorption paths. All these factors contribute to improved sensitivity.

Rather than by measuring the attenuation of a laser beam in the sample, the absorption can often be monitored indirectly with still higher sensitivity. For instance, absorption of a modulated laser beam will produce a sound wave in the sample, because some of the light is converted to heat. This sound wave can be picked up by a microphone (optoacoustic detection). Resonant absorption of laser light by atoms or molecules in a gas discharge can change the ionization probability. The resulting changes in discharge current or voltage across the tube are easily measured (optogalvanic detection). The highest sensitivities at visible and ultraviolet wavelengths have been obtained by monitoring the fluorescence or photoionization of laser-excited atoms or molecules in a gas or molecular beam. Single atoms have been selectively detected in this way.

Intracavity absorption. Very high sensitivity has also been obtained by placing an absorbing sample inside the resonator of a broad-band dye laser without any optical tuning elements. Any absorption lines can be detected as dips in the laser emission spectrum when analyzed by a spectrograph. This intracavity spectroscopy can surpass the sensitivity of a conventional single-pass absorption measurement by a large factor (on the order of 10^5), because the sample is effectively traversed by the light many times, and the competition between many simultaneously oscillating modes enhances the suppression of modes with slightly increased losses.

Multiphoton absorption. It has long been known that atoms can be excited to a higher quantum state by simultaneously absorbing two or more photons which together provide the necessary energy. The probability for an N-photon transition grows initially with the N-th power of the intensity. Only the high intensity of laser light has made it possible to observe two-photon and multiphoton absorption in the optical region. Two-photon spectroscopy permits the study of states with the same parity as the absorbing level, which are not normally reached by single-photon transitions, and it requires photons of less energy, which are sometimes more readily produced. *See* SELECTION RULES (PHYSICS).

Fluorescence spectroscopy. Intense laser light is a very effective means to pump a large fraction of an absorbing species to some excited quantum level. Hence, lasers can greatly increase the sensitivity of such classical spectroscopic methods as fluorescence spectroscopy, optical pumping, level-crossing spectroscopy, or double-resonance spectroscopy. Moreover, lasers make it possible to apply these techniques to atomic and molecular transitions at wavelengths where intense spectral lamps are not available, and stepwise excitation permits studies even of highly excited states, including autoionizing levels and very high Rydberg states of atoms.

Studies of the line shape of fluorescent emission at high intensities have permitted interesting tests of the predictions of quantum-electrodynamic theory. *See* QUANTUM ELECTRODYNAMICS.

Raman spectroscopy. Lasers have revolutionized Raman spectroscopy, the observation of scattered light at wavelengths other than that of the exciting light. The high intensity of laser light has greatly increased the sensitivity of this form of two-photon spectroscopy, where the energy difference between incident photon and scattered photon corresponds to a resonant transition between states of equal parity. Tunable lasers have enhanced the sensitivity, resolution, and versatility of Raman spectroscopy even more, by making it possible to excite close to some intermediate resonant transition (resonance Raman spectroscopy), or to observe stimulated Raman scattering with the help of a second, tunable probe laser beam. Polarization anisotropies associated with Raman transitions provide additional information and can further increase the detection sensitivity. *See* RAMAN EFFECT.

Frequency mixing. The high intensity of laser light has also made possible a new class of spectroscopic methods which rely on nonlinear frequency mixing in the sample. For instance, if a sample is irradiated simultaneously by two strong laser beams of frequencies ω_1 and ω_2, the nonlinear response of the driven dipoles leads to the generation of coherent light at a new frequency, $2\omega_1-\omega_2$. This frequency mixing can be described in terms of a third-order nonlinear susceptibility of the sample.

If the frequency ω_2 is tuned to the (redshifted) Stokes line of a Raman transition, excited by ω_1, a resonance is observed in the intensity of the new wave at the (blueshifted) anti-Stokes frequency. (A

Stokes line is a transition close to a Raman transition at a frequency shifted below the Raman resonance by a quantum of vibration of a molecule. On the contrary, if the frequency is shifted above the Raman resonance by the same quantum of vibration, one has the anti-Stokes line.) This type of frequency mixing is sometimes referred to as coherent anti-Stokes Raman spectroscopy (CARS). It can be used to study Raman transitions even in the presence of strong incoherent background radiation, such as in the observation of flames.

High-resolution laser spectroscopy. Lasers have led to particularly noteworthy progress in the field of very high resolution spectroscopy. They have become powerful tools to investigate the structure of atoms, molecules, and ions. They can be used to study fine and hyperfine splittings, Zeeman and Stark splittings, light shifts, collision broadening, collision shifts, and other attributes of spectral lines. Moreover, lasers make it possible to measure the wavelengths of spectral lines with unprecedented accuracy. The laser wavelength can be locked to an atomic or molecular transition, providing an accurate standard length or frequency. Such lasers have become important tools for precision metrology, making possible accurate measurements of fundamental constants and stringent tests of basic physics laws. *See* RYDBERG CONSTANT.

Doppler-free spectroscopy. The sharpest spectral lines are generally those in atoms or molecules which are relatively free and undisturbed. Such molecules, however, are almost inevitably moving with high thermal velocity. Molecules which are moving toward an observer appear to emit or absorb light at higher frequencies than molecules at rest, and molecules moving away appear to absorb at lower frequencies. In a gas, with molecules moving at random in all directions, the lines appear blurred, with typical Doppler widths on the order of 1 part in 1,000,000. To take advantage of the very narrow instrumental linewidth of laser sources, it is necessary to overcome this Doppler broadening of spectral lines. *See* DOPPLER EFFECT.

Saturation spectroscopy. Laser light can easily be intense enough to partly saturate the absorption of a spectral line. That is, those molecules which have absorbed a light quantum are temporarily removed from the initial state. The absorption from that level is reduced or saturated, at least until some relaxation process can replenish the supply of absorbing molecules. In a gas, a monochromatic laser beam will resonantly interact only with those molecules which have the right axial velocity to be Doppler-shifted into resonance. The resulting velocity-selective saturation can be used for Doppler-free spectroscopy.

In 1970 C. Borde and T. W. Hänsch introduced independently a now commonly used form of saturation spectroscopy which achieves good sensitivity with samples outside the laser resonator and which is particularly well suited for use with broadly tunable laser sources. As in **Fig. 1**, the output of a tunable laser is divided by a beam splitter into a stronger saturating beam and a weaker probe beam that are traversing an absorbing gas sample along the same path but in nearly opposite directions. When the saturating beam is on, it bleaches a path through the cell; that is, it depletes those molecules which are Doppler-shifted into resonance, and a stronger probe signal is received at the detector. As the saturating beam is alternately stopped and transmitted by a chopper, the probe signal is modulated. However, that happens only when both beams interact with the same molecules, and those can only be molecules which are standing still or at most moving transversely. Thus the method picks out those molecules which have near-zero component of velocity along the laser beams and ignores others. **Figure 2** illustrates the power of the technique by comparing a Doppler-broadened absorption profile of the well-known Balmer-alpha line of atomic hydrogen with one of the first saturation spectra, recorded with a pulsed dye laser in a glow discharge.

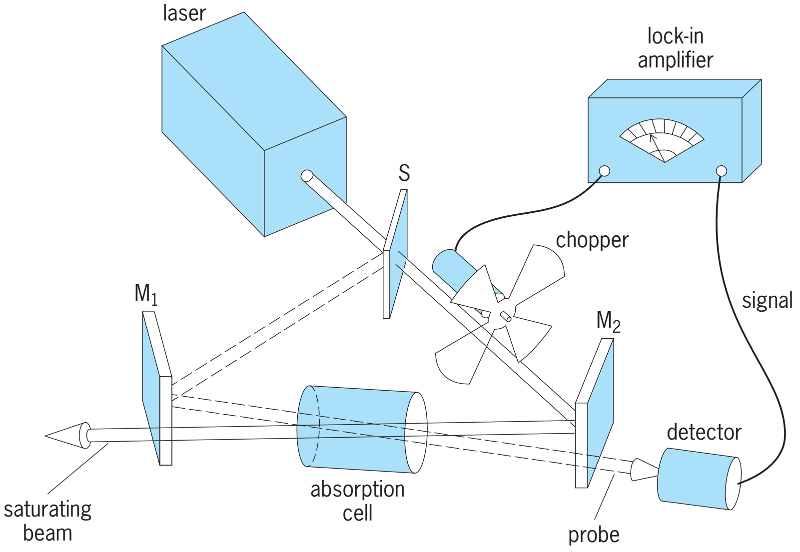


Fig. 1. Scheme of laser saturation spectrometer; S indicates the beam splitter and M_1 and M_2 are mirrors.

The described methods work well only if the sample has noticeable absorption and if the laser is strong enough to excite a substantial fraction of the resonant molecules. Higher sensitivity can sometimes be achieved by observing the absorption of laser light indirectly via the emitted fluorescence or via acoustooptic or optogalvanic detection. In this case, it is advantageous to monitor the nonlinear interaction of two counterpropagating laser beams via intermodulation; that is, the two laser beams are chopped at two frequencies f_1 and f_2, and the spectrum is recorded as a modulation in the signal at the sum or difference frequency, $f_1 + f_2$ or $f_1 - f_2$.

Polarization spectroscopy. There are also purely optical methods, which can be considerably more sensitive than the older saturated absorption method by suppressing the fluctuating background of probe light on which the signal has to be detected. These include, in particular, the technique of polarization spectroscopy introduced in 1976 by C. Wieman and T. W. Hänsch. A polarization spectrometer looks similar to a saturation spectrometer, but takes advantage of the fact that small changes in light

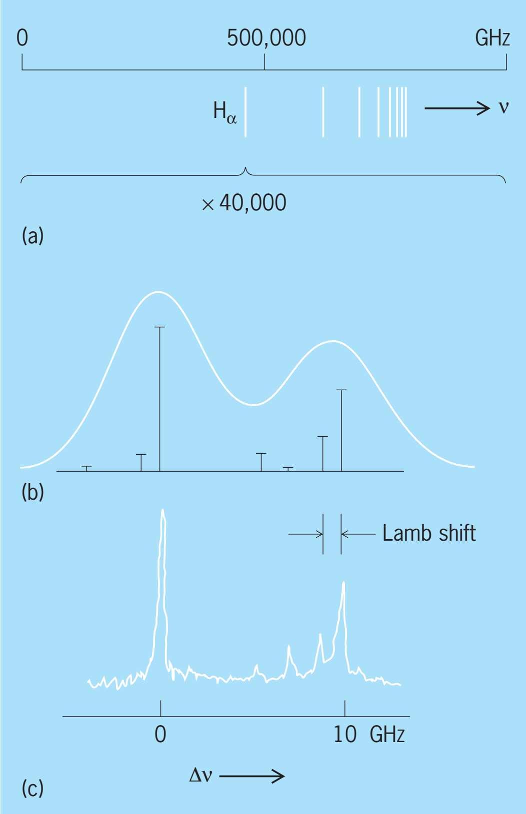


Fig. 2. **Laser spectroscopy of atomic hydrogen. (*a*) Balmer series. (*b*) Doppler-broadened absorption profile (300 K) of the Balmer-alpha line with theoretical fine structure. (*c*) Saturation spectrum of Balmer-alpha line with resolved 2*S*-2*P* Lamb shift.**

polarization can be detected more easily than changes in light intensity. The probe beam "sees" the sample placed between nearly crossed linear polarizers so that only very little light arrives at the photodetector. The saturating beam is made circularly polarized by a birefringent plate. Alternatively, a linearly polarized beam is used with its polarization axis rotated at 45°. *See* POLARIZED LIGHT.

Normally, in a gas, molecules have their rotation axes distributed at random in all directions, but the probability for absorbing polarized light depends on the molecular orientation. Thus the saturating beam depletes preferentially molecules with a particular orientation, leaving the remaining ones polarized. The polarized ones can then be detected with high sensitivity because they can change the polarization of the probe beam. The probe acquires a component that can pass through the crossed polarizer into the detector, but again this happens only near the center of a Doppler-broadened line where both beams are interacting with the same molecules.

Polarization spectroscopy makes it possible to observe fewer molecules with lower light intensity so that external causes of line broadening and shifts are more easily avoided. A promising related technique, saturated interference spectroscopy, can work even in spectral regions where good polarizers are not available.

Doppler-free two-photon spectroscopy. There is a completely different approach to Doppler-free laser spectroscopy which does not rely on velocity selection. However, this approach works only for two-photon or multiphoton spectroscopy, where the frequencies and directions of the exciting photons can be chosen so that the momenta of the absorbed photons add to zero.

Doppler broadening in two-photon excitation can be eliminated to first order simply by reflecting the output of an intense monochromatic laser back onto itself and placing a gas sample in the resulting standing-wave field, so that the molecules can absorb two photons of equal energy coming from opposite directions. From a moving molecule, one beam will appear Doppler-shifted toward the blue, and the other will appear shifted toward the red by an equal amount. The sum frequency is hence constant, independent of the molecular velocity. If the number of excited molecules is observed during a laser scan, a sharp resonance appears on a low Doppler-broadened background, produced by each traveling wave separately. The Doppler-free signal is strongly enhanced because all molecules, regardless of their velocity, contribute.

Laser cooling of trapped ions and atoms. Neutral atoms and trapped ions can be cooled to temperatures on the order of 10 microkelvins with laser light. At those temperatures, the atoms have velocity spreads on the order of 1 cm/s. Both the first- and second-order Doppler shifts of the particles are greatly reduced, and the available measurement time is greatly increased. Laser cooling techniques are broadly applicable to spectroscopic measurements where the signal appears as a small frequency shift. Time standards, tests of mass anisotropy, nonlinearity in quantum mechanics, time reversal invariance, and charge neutrality of atoms are examples of work where laser cooling is playing a role. The resolution of any of the described methods is ultimately limited by the natural linewidth of the observed transition. By using almost forbidden transitions in alkali-earth atoms, it should be possible to achieve a resolution of 10^{-17}. *See* ATOMIC CLOCK.

Laser spectroscopy of cooled ions and atoms has also opened up the study of single-ion and few-ion systems, ultracold plasmas, ultracold collisions, and the formation of dilute quantum gases. These new quantum gases can be used to study, with perfect control, many solid-state phenomena such as superconductivity or electron transport in metals. Furthermore, atoms at these temperatures can be manipulated with potentials created with integrated electronic circuits, opening the way to the development of a totally new generation of quantum devices. Thereby, one could exploit the quantum correlations between the ultracold atoms to realize quantum logic circuits or quantum-limited sensors with unprecedented capabilities. *See* LASER COOLING.

Coherent transients. Nonlinear optical processes of a coherent nature can be observed in the time domain, and are capable of yielding detailed dynamic information about relaxation processes as well as Doppler-free spectra. Coherent transient phenomena, such as optical nutation, free induction decay, and photon echoes arise in a sample excited by a sequence of resonant laser pulses. These phenomena are completely analogous to effects in pulsed NMR spectroscopy. Each transient effect may be used to select and examine a specific dephasing process. The laser field prepares the sample by placing the transition levels in quantum-mechanical superposition, and the resulting dipoles radiate in accordance with the coupled Schrödinger-Maxwell wave equations. The detection of the emitted radiation then yields information about the environment of the oscillators.

Picosecond and femtosecond spectroscopy. Extremely short pulses of light are used to examine events such as vibrational and rotational relaxation of molecules, the thermalization of photo-excited electron plasmas in a semiconductor, and nonthermal processes such as desorption of adsorbates from surfaces. Subpicosecond snapshots of chemical reactions, studies of the primary photochemical behavior of retinal in rhodopsin, and ultrafast electrical pulses are now possible with the short-pulse technology. Most experiments use some variation of the pump-probe technique, where the first optical pulse is used to trigger an event with the absorption of one or more photons. A second pulse optically delayed with respect to the first pulse then probes the system spectroscopically. Coherent transient techniques such as photon echoes or four-wave mixing are also applicable in time domain spectroscopy. The short laser pulses can also be used to generate very short bursts of x-rays, phonons, electrons, and electrical pulses. A very recent development is the achievement of attosecond (10^{-18} s) optical pulses which are capable of directly monitoring electron orbitals in atoms. *See* LASER PHOTOCHEMISTRY.

Optical frequency synthesizers. One of the most recent breakthroughs in laser spectroscopy has been the development of the optical frequency synthesizer, which has completely revolutionized the world of time and frequency standards, allowing the connection of the optical frequency domain with the radio frequency domain. Optical frequency synthsizers are the most precise measurement tool now available, with an inherent stability that has been tested beyond 10^{-18}.

An optical frequency synthesizer is based on a mode-locked pulsed laser capable of emitting pulses that can be made as short as a few femtoseconds. The frequency spectrum of a train of pulses emitted by such a laser is a comb of equally spaced modes (the frequency spacing is just the laser repetition rate) that can span more than one octave. It is possible to compare the two ends of the spectrum by comparing the modes at a given low frequency with the modes at twice that frequency. This can be achieved by frequency-doubling the low-frequency modes in a nonlinear crystal and observing on a fast photodiode the beatnote between the doubled modes and the high-frequency modes. This procedure allows one to know the frequency of each comb mode, but for an offset frequency. The offset frequency can be measured by comparing a suitably chosen mode of the comb with a laser beam stabilized on a stable frequency reference. Indeed, since the modes of the comb can be actively stabilized to the reference, this procedure effectively transfers the stability of the reference to each of the comb modes, which cover the entire visible range, and to the comb frequency spacing, which is in the radio-frequency domain. The intrinsic stability of optical frequency synthesizers is such that at the moment there is a lack of a sufficiently stable reference to take full advantage of the synthesizer. This is one of the main reasons forcing reconsideration of the definition of the second, abandoning the cesium microwave standard in favor of an optical transition either of a laser-cooled atom or of a single trapped ion. *See* FREQUENCY COMB.

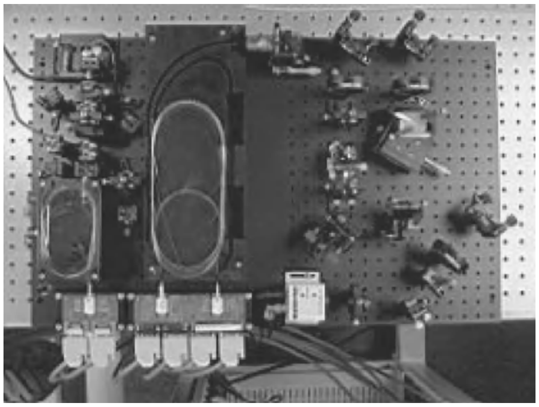


Fig. 3. An optical synthesizer based on a mode-locked fiber laser. The advantage here is that no parts missallign during operation because of slow temperature variations. Thefore this synthesizer can operate continuously for several days.

Other applications. Laser spectroscopy is virtually ubiquitous in natural sciences, but it is also extensively used in many other fields, some as seemingly remote as pollutant monitoring or art restoration. By laser spectroscopy one can readily obtain information on the chemical composition of a given material at a high rate without the need to mechanically touch the material. This is of utter importance in the fabrication of electronic components where spectroscopic techniques can be used to monitor deposition processes while they are being realized.

In analytical chemistry, laser spectroscopy is used for ultrasensitive detection of small concentrations of pollutants, tracer elements, or short-lived intermediate species in chemical reactions. It can be used for the investigation of collision-induced energy-transfer processes or to induce chemical reactions by excitation of one or more of the reactants. Ultrashort laser pulses have led to the birth of real-time femtochemistry where the formation or dissociation of molecules can be observed as it is happening. Laser spectroscopy can also be used for isotope identification and separation.

In biology, laser spectroscopy has been fundamental for the determination of the structure of biological molecules or of energy transfer in DNA, while the extreme time resolution allowed by ultrafast lasers allows the study of fast dynamical processes such as isomerization during photosynthesis.

In medicine, besides the many successful therapeutic applications of lasers, laser spectroscopy is now extensively considered as a diagnostic tool, for example, in breath analysis or in cancer management.

Geophysics and environmental research benefit from laser spectroscopy in the detection of natural-gas emissions or in pollution monitoring. Of particular interest is the LIDAR (light detection and ranging) technique, in which a laser beam is sent into the atmosphere through an expanding telescope and backscattered off a distant object—for example, a volcanic cloud of ashes or a polluted building facade. By measuring the backscattered radiation with spectral and time resolution, it is possible to gain information on the object distance, extension, and composition.

Another example of a laser spectroscopy application is combustion monitoring, which can be performed in situ and in real time, for example, inside a functioning engine; or the measurement of flow velocities and turbulence.

Raman spectroscopy is widely used for the noninvasive determination of the chemical composition of paint in frescoes and paintings as well as stone analysis in sculpture.

Laser spectroscopy can also provide insights about the purity of material, its composition, and the quality of production processes in the fabrication of electronic circuits. Theo W. Hänsch; Francesco S. Cataliotti

Bibliography. D. L. Andrews (ed.), *Applied Laser Spectroscopy*, 1992; W. Demtröder, *Laser Spectroscopy: Basic Concepts and Instrumentation*, 3d ed., 2003; P. Hannaford (ed.), *Femtosecond Laser Spectroscopy*, 2005; T. W. Hänsch and M. Inguscio (eds.), *Frontiers in Laser Spectroscopy*, 1994; E. R. Menzel, *Laser Spectroscopy: Techniques and Applications*, 1995; S. Svanberg, *Atomic and Molecular Spectroscopy: Basic Aspects and Practical Applications*, 4th ed., 2004.

Laser welding

Welding with a laser beam. The primary apparatus used for both commercial application and industrial research and development is the continuous-wave, convectively cooled CO_2 laser with either oscillator/amplifier (gaussian output beam) or unstable resonator (hollows output beam) optics. These lasers are available in output powers ranging from approximately 1000 to 15,000 W. Substantial advances in laser technology made possible the production of fully automated multikilowatt industrial laser systems which can be operated on a continuous production basis. These systems can be used for a variety of development programs and on-line production applications.

Welding of HY-130 steel. A significant metallurgical phenomenon was reported following the laser welding of HY-130 alloy steel. After welding with a 7-kW continuous-wave coaxial-flow CO_2 laser with oscillator-amplifier optics, concurrent increases in the hardness, tensile strength, and impact energy of the welds were noted when compared with base metal properties. Welding was accomplished in the open atmosphere under gas shielding with no preheating, postheating, or filler metal additions. A decrease in the visible inclusion content was noted both on the impact fracture surface (**Fig. 1**) and in metallographic sections. Chemical analysis revealed significant decreases in the oxygen and nitrogen contents of the weld metal as compared to the base metal. It was concluded that during laser welding of this alloy a purification of the fusion zone with respect to inclusions, and perhaps a change to a more favorable inclusion size distribution, occurs.

Single- vs dual-pass welds. Single- and dual-pass laser welds have been made in an alloy of the 80,000-lb/in.^{2} (5.5×10^8 pascals) yield strength class. **Figure 2** shows cross sections of these welds. The welds were formed with a 12-kW beam from a continuous-wave CO_2 laser. Welding speeds of 25 and 30 in./min (1 and 1.3 cm/s) and 60 and 65 in./min (2.5 and 2.8 cm/s) were used for the single- and dual-pass welds, respectively. The welds were evaluated and exhibited excellent overall mechanical properties. Dual-pass welds exhibited a ductile-to-brittle transition temperature below −60°F (−51.1°C). The increased shelf energy was attributed to a reduction in the visible inclusion content of the fusion zone, while transition temperature was shown to be strongly dependent on grain size.

Titanium alloys. Two in-depth studies involving laser welding of Ti-6Al-4V alloy were conducted. In one study, laser welding was compared with electron-beam and plasma-arc welding. Sound welds were made by all processes, and the tensile properties of these welds exceeded base metal properties. The laser and electron-beam welds, because of their lower specific energies for fabrication, were substantially finer-grained and possessed finer substructures than the plasma-arc welds. Due to the finer microstructure of the beam welds, they exhibited lower K_{IC} (fracture toughness in plane strain) values than the plasma-arc welds, which had a coarser structure. The finer structure of the laser and electron-beam welds showed potential for good fatigue crack initiation resistance.

A comparison study of the fatigue behavior of laser and plasma-arc welds was also conducted. It was found that both techniques were capable of making weld specimens where fatigue failures initiated in the base metal; however, there were a number of specimens welded by each process in which fractures initiated at pores in the weld.

Aluminum alloys. While successful weld penetrations have been made in some aluminum alloys, and reasonable-looking beads have been produced, no extensive weld-quality studies or mechanical-property studies have been reported. Most weld penetrations in aluminum alloys show some porosity, but it appears that the potential of welding aluminum does exist.

Industrial applications. A number of multikilowatt laser welding systems are in use for industrial welding and heat-treating applications, including Ford Motor Company, Western Electric Company, Caterpillar Tractor, and General Motors Corporation.

A system which clearly displays the advantages of laser processing is the underbody welding system at

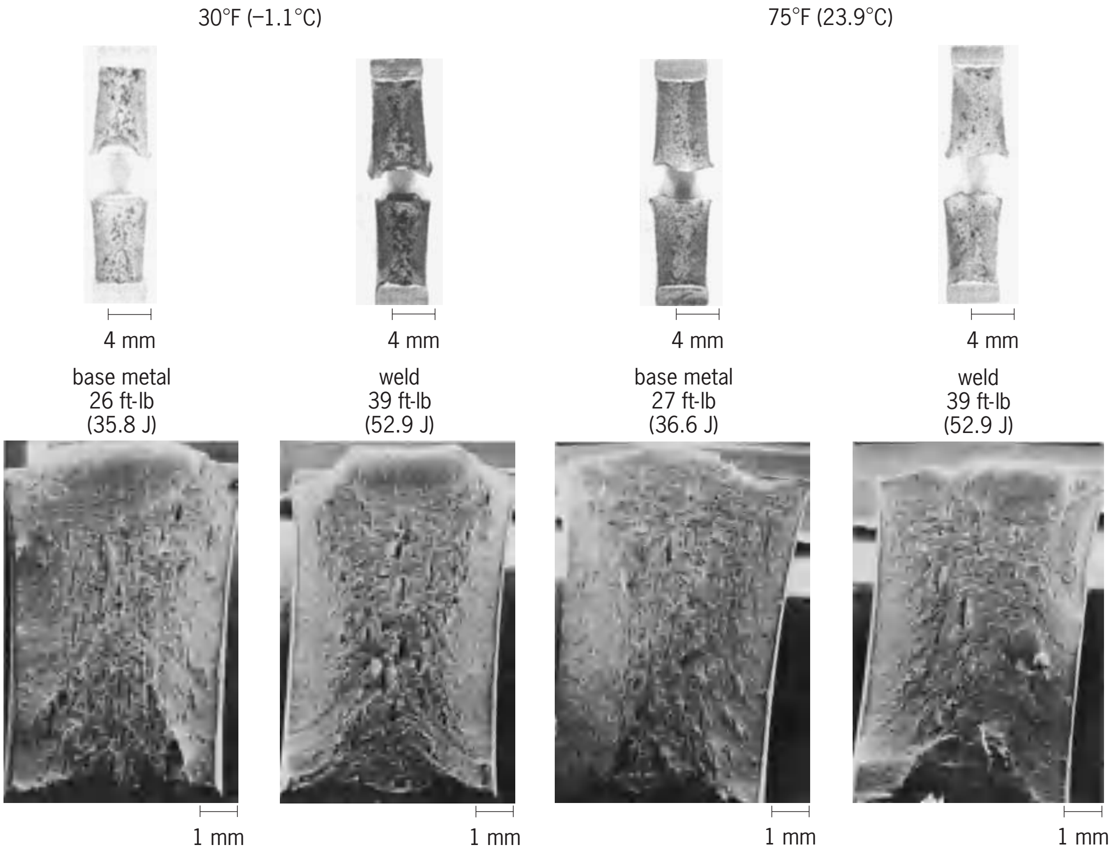


Fig. 1. **A change in the shape and distribution of the inclusions on the fracture surfaces of HY-130 weld and base metal impact specimens is shown. The top row comprises low-magnification macrophotographs of both halves of the fractured specimens. The bottom row comprises higher-magnification scanning electron micrographs of one-half of the fracture. Welds were made at 5 kW and a speed of 45 in./min (114 cm/min). (*From E. M. Breinan and C. M. Banas*, *Fusion zone purification during welding with high power CO_2 lasers*, *Proceedings of the 2d International Symposium of the Japan Welding Society*, *Osaka*, *August 24–28*, *1975*)**

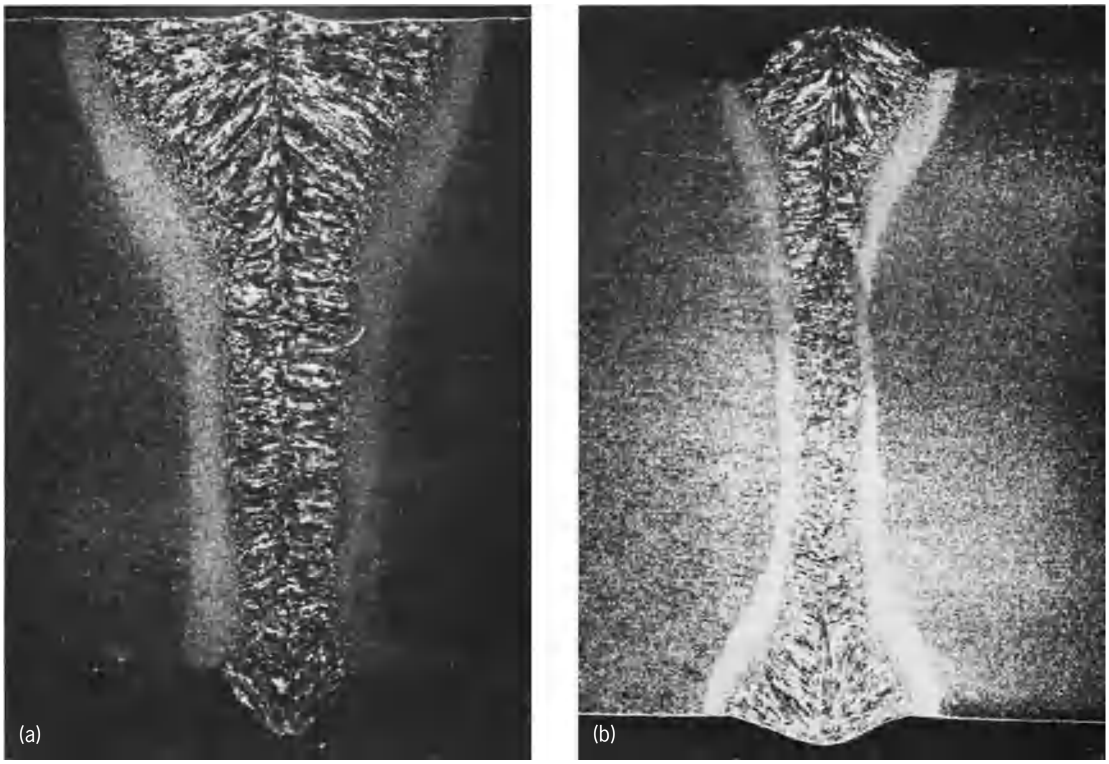
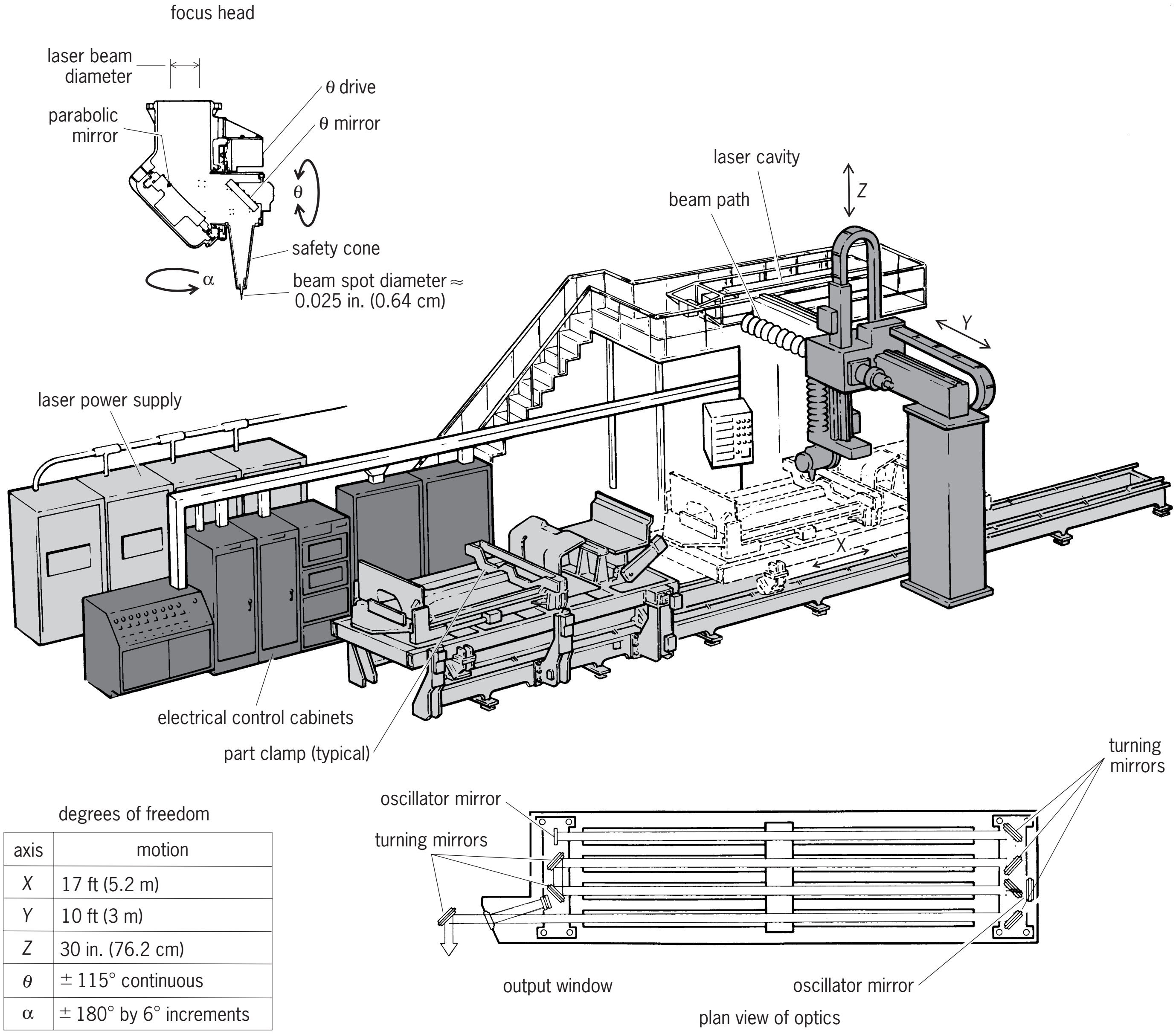


Fig. 2. **Cross sections (0.5-in.- or 1.3-cm-thick plate) of laser welds in X-80 arctic pipeline steel. (*a*) Single-pass weld. (*b*) Dual-pass weld. (*From E. M. Breinan and C. M. Banas, Weld. Res. Counc. Bull. 201*, *pp. 47–57*, *December 1974*)**



degrees of freedom

axis	motion
X	17 ft (5.2 m)
Y	10 ft (3 m)
Z	30 in. (76.2 cm)
θ	± 115° continuous
α	± 180° by 6° increments

Fig. 3. Schematic diagram of 6-kW multiaxis laser welding system at Ford Motor Company; arrows indicate direction of motion along various axes. This is the first complete high-power laser welding system designed to perform a high-production-rate industrial fabrication operation. (*Hamilton Standard Division*, *United Technologies Corp.*)

Ford (**Fig. 3**). As illustrated, a total of five degrees of freedom are achieved in movement of the focused beam relative to the intricately curved underbody panels. Only one of these, the X direction, is achieved by mechanical actuation of the parts; all the rest are achieved by rotation or translation of water-cooled copper mirrors. A unique 90° off-axis parabolic focusing mirror turns the beam 90° as it focuses the beam, thus allowing easy implementation of the α and θ axes. The focused beam creates a continuous lap weld in the thin steel sheet to a depth of 0.06 in. (0.15 cm) and at a speed of 500 in./min (21 cm/s). The continuous weld achieves much greater strength than spot welding techniques, and assures leak-tightness of the joint. At that welding speed, four large panels are joined together into an underbody in a total cycle time of 1 min.

Future applications. As the power available in commercial lasers increases, the scope of applications is also increasing. In addition to assembly-line operations, several opportunities are seen in the area of welding of heavy sections for industrial purposes including shipbuilding, pipeline welding, and general heavy steel fabrication. The use of added filler metal, which has already proved feasible with laser, promises to extend the thickness capability of the laser welding process.

Laser welding has several potential advantages for deep welding tasks. For one thing, high welding speeds are achievable. In addition, the lack of necessity for filler metal or substantial reduction in the amount of filler required, promises economic advantages for the laser process, particularly when welding arctic pipeline steels and in other applications normally requiring large quantities of expensive filler metal. Reductions in edge preparation cost may also be realized due to the elimination of the edge chamfer required, although some machining of the butt edges is usually necessary to achieve a joint fitup which is suitable for laser welding. Another important factor is the ability of a single laser installation to supply a beam to several different welding stations, by appropriate indexing of turning mirrors. This capability, along with the high welding speed, allows high productivity on a laser welding installation.

A final advantage to laser welding is in the quality of the joint, apart from the low cost of welding. For quality-critical applications, such as welding of nuclear reactor components or arctic pipeline steel, the joint quality alone may prove to be an important factor in the selection of laser welding. *See* WELDING AND CUTTING OF MATERIALS. Edward M. Breinan

Bibliography. C. Dawes, *Laser Welding*, 1992.

Lateral line system

A primitive vertebrate sensory system that is present in all larval and adult fishes, in larval amphibians (such as tadpoles), and in some adult amphibians that retain an aquatic lifestyle (including the clawed frog). It is absent from all reptiles, birds, and mammals, even those that are aquatic (such as turtles, dolphins, and whales). The lateral line system consists of 100 or more sensory organs (neuromasts) that are typically arranged in lines on or just under the skin of the head and body. Neuromasts are composed of sensory hair cells, which are also found in the auditory system of all vertebrates. The lateral line system responds to water flowing past the skin surface and uses different flow patterns over the body to form hydrodynamic images of the animal's nearby surroundings, just as the visual system forms visual images of the environment using different light patterns on the retina.

Structure and function. Neuromast receptor organs are generally round or oval and can range from 20 micrometers to over 1000 μm (0.001 to 0.04 in.) in length. A neuromast has a central population of up to hundreds of mechanosensory hair cells that are surrounded by support cells. Hair cells function as directional sensors that convey information to the brain about both the strength and direction of water currents. The ciliary bundle of each hair cell contains one long cilium (kinocilium) and a cluster of shorter cilia (stereocilia), which lead up to the kinocilium in a stepwise fashion—the shortest being farthest away from the kinocilium and the longest closest to the kinocilium (**Fig. 1**). The ciliary bundles of the hair cells are embedded in a gelatinous cupula. Hair cells are activated when water flows past the skin surface, causing the cupula to move, thus causing the cilia to bend. The neural response of each hair cell is proportional to both the degree of cilia displacement and the direction in which the stereocilia are displaced relative to the eccentrically placed kinocilium of each hair cell (Fig. 1).

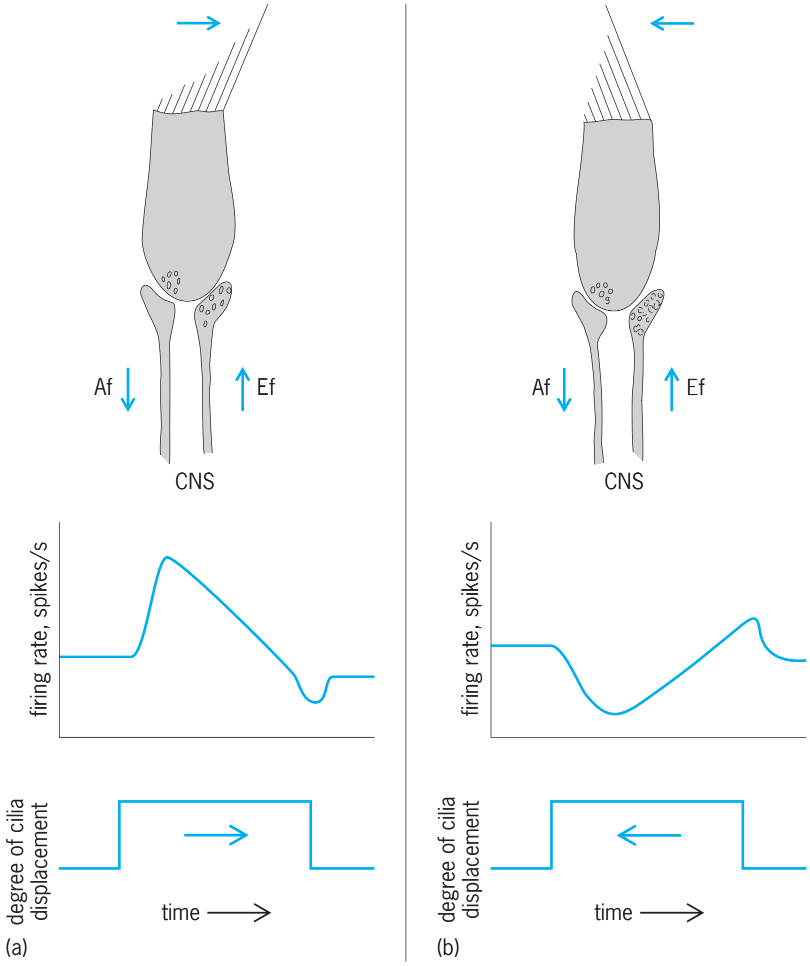


Fig. 1. Hair cells of the lateral line system responding to water flows in two opposing directions. (*a*) Stereocilia displaced toward the elongated kinocilium produce an increase in the firing rate of afferent (Af = afferent, Ef = efferent) fibers carrying information from the hair cell to the central nervous system (CNS). (*b*) Stereocilia displaced away from the kinocilium result in a decrease in neural activity.

Distribution on the head and body. Neuromasts are found just under the skin in fluid-filled canals that communicate with the skin surface through a series of pores (canal neuromasts, found in fishes only), or on the skin surface (superficial neuromasts, found in fishes and amphibians). A prominent canal that often forms a visible line along the trunk of most fishes is probably the origin of the term "lateral line," but in reality the lateral line system includes neuromasts that are distributed all over the head and body of the animal. In fishes, lateral line canals on the head are embedded in bone (bony fishes) or cartilage (cartilaginous sharks, skates and rays), whereas body (trunk) canals are contained in or between scales. Head canals in bony fishes may also vary in width, number of pores, and degree of development. In amphibians, canals are not present. Instead, short rows of up to 10 or so superficial neuromasts form a stitch, and these rows are arranged in lines along the head and trunk of the animal. Canals in fishes and stitch lines in amphibians are distributed in relatively stereotyped patterns (such as above and below the eye, down the cheek, along the lower jaw, across the top of the head, and down the main body) [**Fig. 2**].

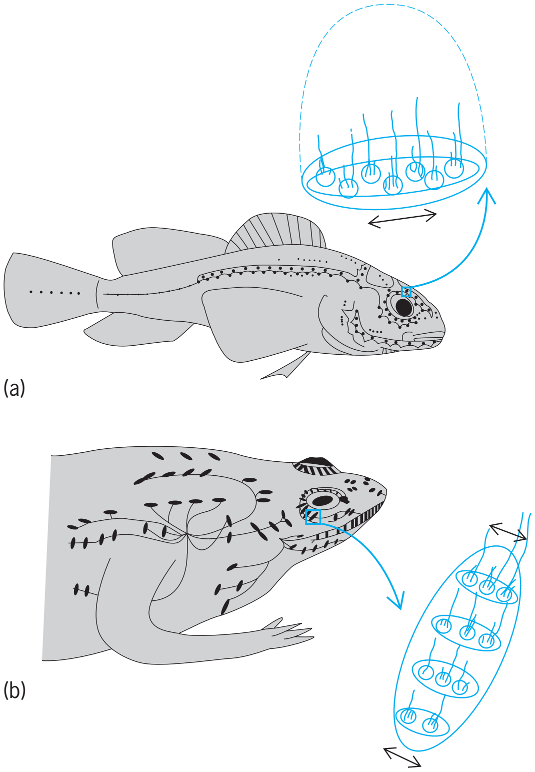


Fig. 2. Lateral line system of a typical bony fish and amphibian. (*a*) In bony fish, neuromasts (black dots) are found on the skin and in pored canals. The hair cells of each neuromast (closeup) are overlain by a dome-shaped gelatinous cupula (dotted line). In amphibians, neuromasts are found only on the skin in stitches (black ovals) that are composed of several neuromasts (closeup).

Superficial neuromasts in fishes are usually smaller (less than 100 μm or 0.004 in. in diameter) and contain fewer hair cells (less than 100) than canal neuromasts. They may occur in rows or singly, sitting flush with the epithelium, in small pits, or on top of papillae or ridges of skin. The number and distribution of superficial neuromasts can also vary considerably among species, ranging from less than 50 distributed in a few (about 10) distinct locations to more than 1000 distributed all over the head and body. *See* TELEOSTEI.

Processing of information. Information from neuromasts is transmitted to the brain by sensory (afferent) nerve fibers, which form five cranial nerves (the lateral line nerves) that terminate in distinct medullary regions of the brainstem (medulla oblongata). Contrary to many earlier published reports, these nerves are distinct from cranial nerves V, VII, IX, and X, so fishes and amphibians have more than the 12 cranial nerves traditionally described in mammals. Distinct regions and pathways in the brain are dedicated to processing information from the lateral line system. These are similar in overall organization, and in proximity to regions of the brain dedicated to processing information from two other closely allied sensory systems, the auditory system and the electrosensory system. Information is also carried from the brain to the sense organs by efferent nerve fibers, which can modulate the sensitivity of the organs to certain stimuli (for example, to reduce sensitivity when water flows are produced by the animal's own movements). *See* AMPHIBIA.

Behavioral significance. The lateral line system is thought to have a function that is separate from, but intermediate between, touch and hearing, and is best described as a sense of touch-at-a-distance. The lateral line system responds to both unidirectional (DC) and oscillatory (AC) water flows that change directions up to 200 times (cycles) per second. In general, large-scale water movements such as oceanic currents, tides, and river flows that are strong enough to carry a fish with them are not by themselves very effective lateral line stimuli. Smaller-scale movements, such as those produced by a slowly moving (less than 8 cm or 0.25 ft/s) stream or by nearby animals (less than one or two body lengths away), can be effective lateral line stimuli. Living organisms can produce water flows in different ways (such as inhalant and exhalant respiratory currents produced by bivalve mollusks, the turbulent wakes left by swimming fishes, and the fine antennal movements of tiny planktonic invertebrates).

Fishes can use different types of water flows to form hydrodynamic images of their surroundings. They can form images passively by remaining still and simply detecting the water currents created by other moving animals, or by detecting current distortions or turbulent wakes created by a stationary obstacle in moving water. Alternatively, fishes can actively form images by swimming past a stationary obstacle and then detecting the distortions in their own self-generated flows due to the presence of the obstacle. In some fishes, such as blind cavefishes, thousands of superficial neuromasts that are over the entire body surface are specialized for active imaging. These fishes are capable of determining very fine spatial details (less than 2 mm or 0.08 in.) of their surroundings.

Fishes and amphibians can also use their lateral line system to orient themselves relative to a water current (rheotaxis), hold a stationary position in a stream, capture prey, avoid predators, and communicate with intraspecifics (**Fig. 3**). Many stream-dwelling fishes (such as trout and salmon) show rheotactic and station-holding behaviors by orienting their bodies upstream and holding positions behind stationary rocks or boulders. These behaviors are important for the upstream spawning migrations of these fishes and for capturing prey that are being carried downstream.

Many different species use their lateral line system to detect prey. Some fishes (such as the topminnow) and amphibians (such as the clawed toad) have specialized canals or receptor organs on their dorsal surface for detecting surface waves produced by struggling insects. Others (such as the Atlantic stingray) appear to have specializations on their ventral surface for detecting benthic prey. Although the functions of the lateral line system in deep-sea fishes have not yet been determined, lateral line morphology in these fishes is frequently very distinctive, with widened canals covering nearly the entire surface

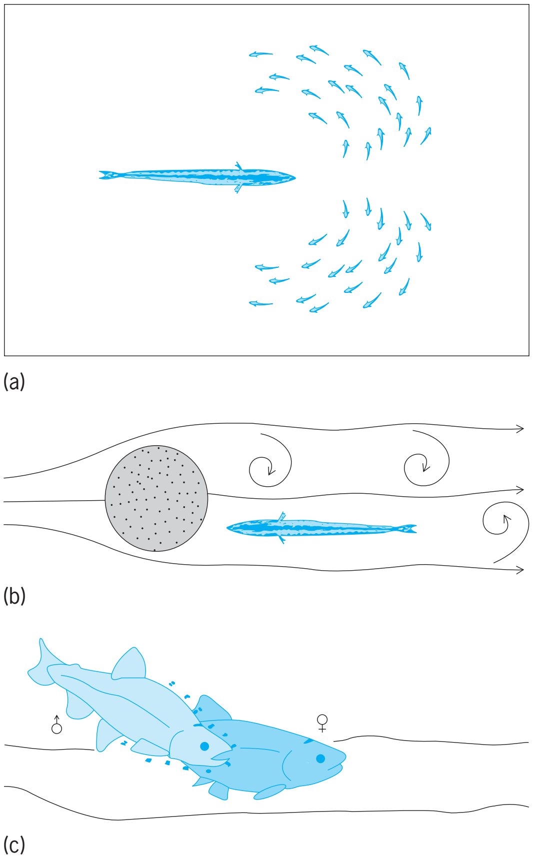


Fig. 3. Uses of lateral line system. (*a*) Rapid schooling maneuver made by a school of herring to escape a rapidly approaching predator. (*b*) Brook trout orienting upstream and holding station behind a large boulder. (*c*) Male salmon quivering next to a female salmon in preparation for mating.

area of the head (as in Poromitra) and prominent superficial neuromasts that sit on the ends of papillae, which elevate the receptors above the skin surface (as seen in the deep-sea anglerfish).

Lateral line specializations are also present in other groups of fishes, including herring, which have a unique connection between a lateral line canal on the head and an air cavity inside the head (near the ear) that functions as an accessory auditory structure. Herring also have extensively branched tubules that connect canals to the outer skin surface and that terminate in many pores. These specializations may allow these fishes to make rapid and coordinated schooling maneuvers to avoid predation. The body movements of nearby fishes can be detected by the lateral line system, and fishes can use this information to help match their swimming velocities and maintain constant distances with adjacent fishes in a school. Body vibrations exchanged between males and females during elaborate courtship dances may serve as another form of intraspecific communication. In land-locked salmon, these vibrational signals are detected by the lateral line system, and they are used to help synchronize the release of gametes (sperm and eggs) so that external fertilization can occur. *See* ELASMOBRANCHII; NERVOUS SYSTEM (VERTEBRATE).

Sheryl Coombs; Jacqueline F. Webb

Bibliography. J. Atema et al. (eds.), *Sensory Biology of Aquatic Animals*, Springer-Verlag, New York, 1988; S. Coombs, P. Görner, and H. Münz (eds.), *The Mechanosensory Lateral Line: Neurobiology and Evolution*, Springer-Verlag, New York, 1989; J. Montgomery, S. Coombs, and M. Halstead, Biology of the mechanosensory lateral line in fishes, *Rev. Fish Biol. Fisheries*, 5:399–416, 1995.

Lateral meristem

Strips or cylinders of dividing cells located parallel to the long axis of the organ in which they occur. Radial enlargement of the cells derived from these meristems increases the diameter of the organ. The lateral meristem is concerned with secondary growth in the sense that its meristematic activity adds cells to the primary body which was derived from the apical meristems. *See* APICAL MERISTEM.

Vascular cambium. The vascular cambium produces secondary xylem on its inner face and secondary phloem on its outer face. In a woody plant the cambium occurs as a continuous sheath, or hollow cylinder, separating xylem and phloem. In stems with separate vascular bundles, fascicular cambium arises within the bundles as strips from the residual procambial cells lying between the primary xylem and primary phloem. In plants with uninterrupted cylinders of secondary vascular tissues, interfascicular cambiums often arise in the ray parenchyma between the bundles in lateral connection with the fascicular cambial strips within the vascular bundles; thus the alternating strips of fascicular and interfascicular cambium form a complete cambial cylinder (**Fig. 1**). The vascular cambium in woody roots arises as strips on the inner faces of the primary phloem groups, and it becomes a laterally continuous sheath when the pericycle cells outside the primary xylem ridges become transformed into cambial initials. *See* PHLOEM; ROOT (BOTANY); XYLEM.

The vascular cambium has two kinds of initial cells: ray and fusiform, or spindle-shaped (**Fig. 2**). Because the immediate derivatives of these cambial cells are quite similar to the initials and may undergo further division, these derivatives and the initials together are said to constitute the cambial zone. The zone may be several cell layers in radial dimension. The composition of the cambium itself, even in a given plant, is not static, for initials may be lost or gained or markedly changed in form. The cambium keeps pace circumferentially with the increasing girth of an organ by self-multiplication of the fusiform initials and by the origin of entirely new groups of ray initials.

Fusiform initials. These are vertically elongate cells, brick-shaped in transectional view, with tapering ends, and with an extensive vacuolar system. These initials give rise to all cells in the vertical system of secondary xylem and phloem. As seen in tangential views, storied cambium has an orderly arrangement in horizontal tiers of uniformly short fusiform initials whose ends slightly overlap those of

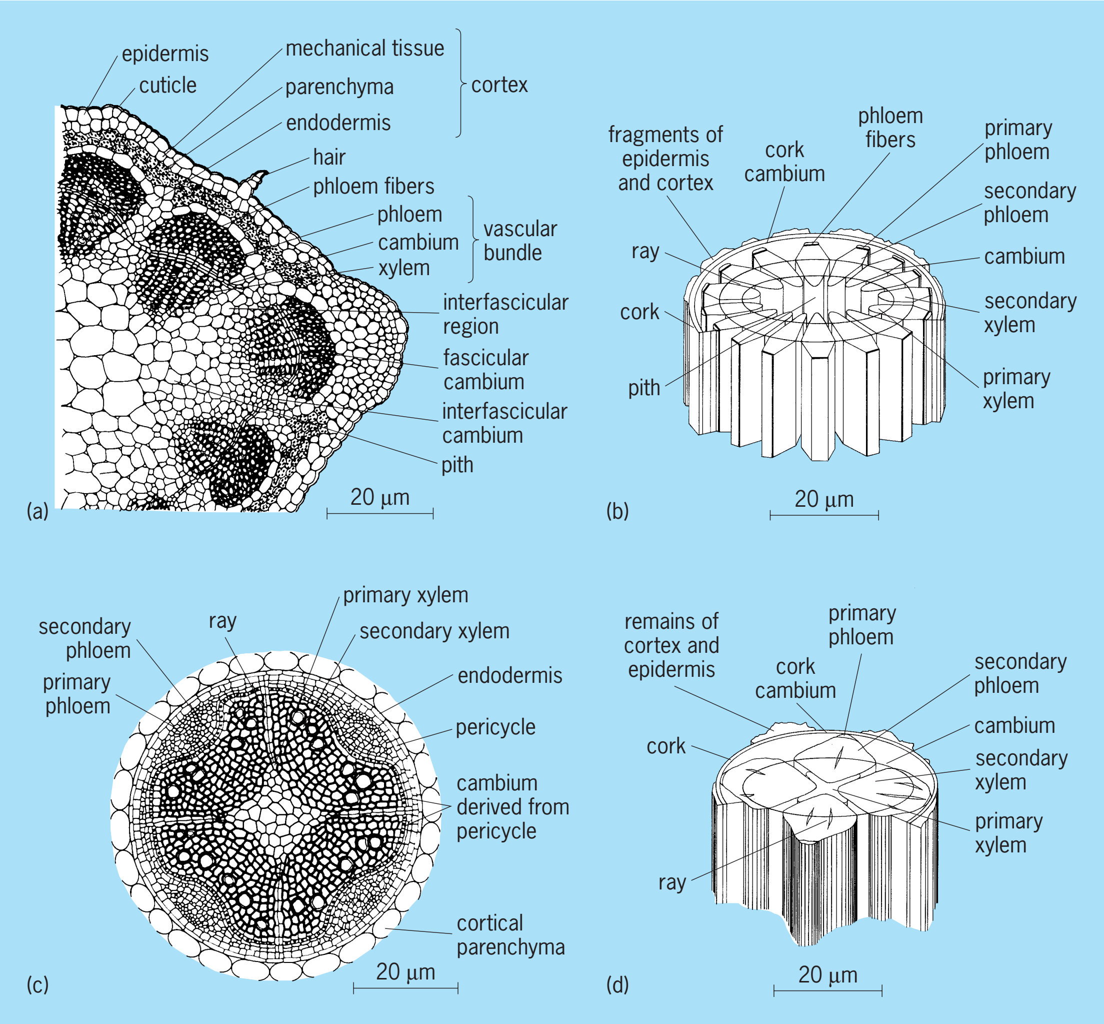


Fig. 1. Diagrams of stem and root. (*a*) Part of a cross section of an alfalfa stem. (*b*) Three-dimensional diagram of a dicotyledonous stem, showing the results of cambial activity. (*c*) Cross section of the vascular cylinder of a dicotyledonous root at level where cambium has become continuous layer and has produced considerable amounts of secondary tissues. (*d*) Three-dimensional diagram of part of root with considerable secondary thickening. (*After G. M. Smith et al.*, *Textbook of General Botany*, *5th ed.*, *Macmillan*, *1953*)

initials in juxtaposed tiers. The more primitive and less orderly nonstoried cambium lacks obvious tiers, and has longer initials of variable length whose ends greatly overlap those of other initials. The radial walls of fusiform initials are relatively thick (with many depressed primary pit fields) in the dormant season, but become much thinner during part of the growing season when the cells are actively dividing. Divisions in the tangential (periclinal) plane produce secondary tissues. Multiplication of initials—which may occur in the dormant season—results from radial (anticlinal) divisions. The separating walls in these divisions may be highly oblique to almost transverse in long initials, or approximately vertical in short initials. The derivatives of oblique divisions elongate by apical intrusive growth.

Ray initials. These cells are nearly isodiametric in form and, though much shorter vertically, often have a wider radial dimension than do fusiform initials. Ray initials divide periclinally in producing new phloem and xylem ray cells. They may also divide

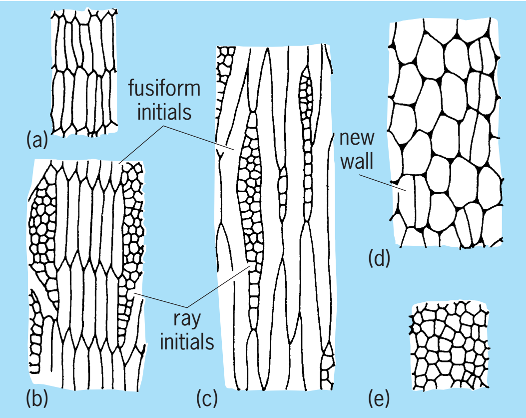


Fig. 2. Tangential views of cambium. (*a*) Root, *Bougainvillea spectabilis*, cork cambium with but one type of initial which is elongate polygonal in shape. (*b*) Stem, *Robinia pseudoacacia*, storied vascular cambium with ray and short fusiform initials. (c) Stem, *Juglans regia*, nonstoried vascular cambium with ray and long fusiform initials. (*d*) Stem, *Yucca gloriosa* (a monocotyledon), unusual vascular cambium with but one type of initial. (e) Stem, *Lindera benzoin*, cork cambium with but one type of initial, nearly isodiametric in shape.

anticlinally or transversely and thus add to the number of initials in a group. An entirely new group of ray initials may have its origin in a segment cut off from the end of a fusiform initial or through subdivision of an entire fusiform initial.

Cork cambium. This cambium (phellogen) produces cork (phellem) on its outer face and sometimes phelloderm (parenchyma) on its inner face. The initials, unlike those in the vascular cambium, are of a single type. They are short and appear rectangular and flattened in transectional and radial views, and rectangular to polygonal in tangential aspect. Cork and phelloderm are formed by periclinal divisions. The increase in the circumference of the cambium occurs by anticlinal divisions. The phellogen initials in certain monocotyledons divide only a few times and are then transformed into cork cells themselves. *See* PERIDERM.

Other cambiums. The initials responsible for production of secondary tissues in certain woody monocotyledons arise in fundamental parenchyma outside the primary bundles. The initials are short and are variable in shape in tangential aspect. They divide in several vertical planes in producing secondary vascular bundles, but strictly periclinally in forming secondary parenchyma.

Some vascular cambiums differ from the ordinary type chiefly in their unorthodox activity. For example, several concentrically oriented cambiums are simultaneously active in the garden beet. Cambiums in some species may remain active for a time and then become transformed into vascular tissue. New cambium then originates outside the secondary phloem. Some cambiums add both xylem and phloem to the inner side. *See* PLANT TISSUE SYSTEMS.

Vernon I. Cheadle

Laterite

The name given by F. Buchanan in 1807 to the iron-rich weathering product of basalt in southern India. The term is now used in a compositional sense for weathering products composed principally of the oxides and hydrous oxides of iron, aluminum, titanium, and manganese. Iron-rich or ferruginous laterite is largely hematite, Fe_2O_3, and goethite, $HFeO_2$, and may be an ore of iron and nickel (Cuba, New Caledonia). Aluminous laterite is composed of gibbsite and boehmite, and is the principal ore of aluminum. Clay minerals of the kaolin group are typically associated with, and are genetically related to, laterite. Laterites range from soft, earthy, porous material to hard, dense rock.

Concretionary forms of varying size and shape commonly are developed. The color depends on the content of iron oxides and ranges from white to dark red or brown, commonly variegated.

Origin. Laterite is formed by weathering under conditions that lead to the removal of silica, alkalies, and alkaline earths. The resulting concentrations of iron and aluminum oxides sharply differentiate laterization from temperate-climate weathering in which the end product is largely clay minerals (hydrous aluminum silicates). Early workers from temperate regions considered lateritization as profound leaching (desilication) in a stage beyond ordinary kaolinization. Studies in tropical regions, however, show that weathering of alkaline silicates may yield gibbsite directly without passing through an intermediate clay stage.

Investigations in many parts of the world stress certain genetic factors. A tropical to subtropical climate with high temperature and abundant rainfall, seasonal or at least with periods of marked dryness, is fundamental. Relief sufficient to ensure good drainage is requisite: lateritic soils are permeable and do not erode by sheet wash like clay or shale. Aluminous laterite forms above the water table and may grade to clay in depth. It may be found on hills and on well-drained slopes, but the residual deposits in adjacent valleys usually are kaolin.

Parent materials. The parent material controls or greatly influences the composition of laterites, which may be developed from a variety of igneous, metamorphic, and sedimentary rocks. Iron-rich rocks (peridotite) yield iron ore; aluminous rocks (syenite) produce bauxite; whereas andesite or basalt give intermediate products. Commonly textural and structural features of the parent material are preserved and the more resistant insoluble minerals remain.

Mature lateritic soils lack fertility for most systems of agriculture. Savannas or parklike grasslands are typical on laterite. Clay, not laterite, is found beneath rainforests and jungle vegetation. *See* BAUXITE; CLAY MINERALS; KAOLINITE; WEATHERING PROCESSES.

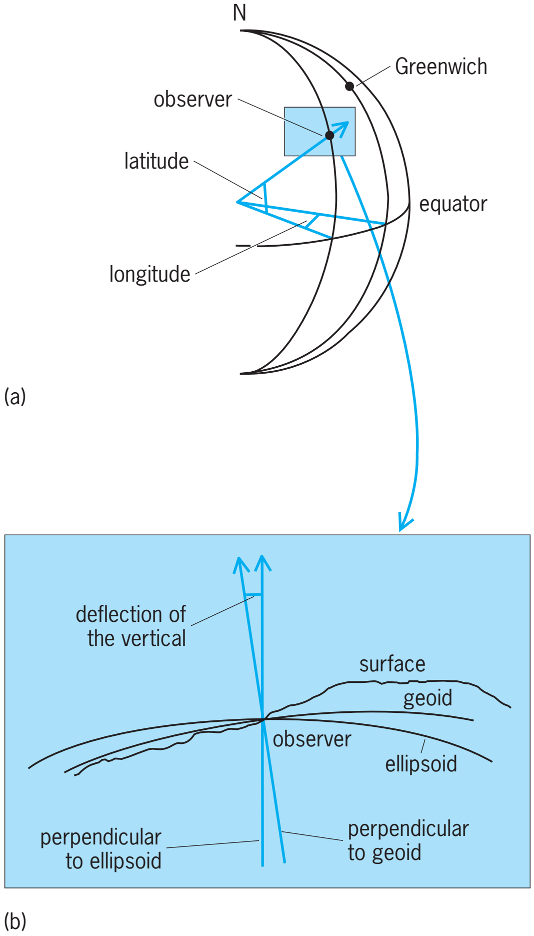
Samuel S. Goldich

Latitude and longitude

The latitude of a location specifies the angle between an imaginary line directed generally toward the center of the Earth and the Equator. The longitude measures the angle between the meridian (the plane defined by the Earth's axis and this local reference direction) and the plane of the Greenwich meridian (**illus.** *a*).

Astronomical coordinates. Astronomical (or astronomic) latitude and longitude use the direction of gravity for the reference direction. This direction, known as the astronomical vertical, is perpendicular to the equipotential surface of the Earth's gravitational field at the location of the observer. *See* ASTRONOMICAL COORDINATE SYSTEMS.

Geodetic coordinates. A particular geopotential surface approximating mean sea level in the open ocean is called the geoid. A mathematical surface in the form of an oblate ellipsoid may be constructed to approximate the geoid. The direction perpendicular to this reference ellipsoid at the observer's location is used as the reference direction in defining geodetic latitude and longitude. A geodetic datum is defined by its reference ellipsoid, the adopted coordinates of a reference station, and the azimuth of a reference



(a)

(b)

Latitude and longitude. (*a*) Angles formed by reference direction, defining latitude and longitude. (*b*) Detail, showing angle between reference directions in astronomical and geodetic coordinates, the deflection of the vertical.

line. Many geodetic datums are in use. *See* MATHEMATICAL GEOGRAPHY.

Geocentric coordinates. Geocentric latitude and longitude are defined by a reference direction which passes precisely through the center of mass of the Earth. These coordinates are determined mathematically from the geodetic latitude and longitude, assuming a fixed relationship between the center of the geodetic datum and the center of mass and knowing the mathematical shape of the ellipsoid. *See* GEODESY.

Determination. Astronomical latitude is determined by observing the altitude on the meridian of a celestial object whose declination is known. Astronomical longitude is identical with the difference between Universal Time and local mean time. Local mean time is determined by observing the time of meridian transits of celestial objects with known positions. The direction of gravity is established with the aid of a spirit level, a liquid surface, or observation of the horizon corrected for the elevation of the observer. It varies locally, and is generally not directed toward the center of the Earth. *See* TIME.

Geodetic latitude and longitude are determined by referring precise measurements of distance and direction, made with geodetic instruments on the surface of the Earth, to the datum. Geodetic coordinates are used for location of sites, while astronomical latitude and longitude are used to determine the angular orientation of celestial objects with respect to observers on the Earth. The difference between astronomical and geodetic coordinates is called the deflection of the vertical (illus. *b*), and commonly amounts to some seconds of arc, occasionally reaching a minute of arc. *See* EARTH, GRAVITY FIELD OF; GEODESY.

Dennis D. McCarthy

Lattice (mathematics)

Lattice theory deals with properties of order and inclusion, much as group theory treats symmetry. As a generalization of boolean algebra, lattice theory was first applied around 1900 by R. Dedekind to algebraic number theory; however, its recognition as a major branch of mathematics, unifying various aspects of algebra, geometry, and functional analysis, as well as of set theory, logic, and probability (to which boolean algebra had already been applied), dates from the years 1933–1938. *See* BOOLEAN ALGEBRA; GROUP THEORY; SET THEORY.

The most basic concept of lattice theory is that of a partial ordering of a set S of elements $x, y, z, \ldots$. By this is meant a binary relation, usually denoted $\leqq$ (or $\geqq$), with the following properties:

$$\text{(P1)} \qquad x \leqq x \text{ for all } x \in S$$

$$\text{(P2)} \qquad \text{If } x \leqq y \text{ and } y \leqq x, \text{ then } x = y$$

$$\text{(P3)} \qquad \text{If } x \leqq y \text{ and } y \leqq z, \text{ then } x \leqq z$$

If $\leqq$ is any partial ordering of S, then its converse or dual $\geqq$, defined by statement (1), is also a partial

$$x \leqq y \text{ if and only if } y \leqq x \tag{1}$$

ordering of S. This easily verified fact provides a fundamental duality principle, which is useful in many connections.

Suppose, for example, the join $x \cup y$ of two elements x and y of a partially ordered set S is defined by conditions (2*a*) and (2*b*). (It is easily shown that

$$x \leqq x \cup y \qquad y \leqq x \cup y \tag{2a}$$

$$\text{If } x \leqq z \text{ and } y \leqq z\text{, then } x \cup y \leqq z \tag{2b}$$

there is at most one such $x \cup y$.) Then the duality principle suggests defining the meet $x \cap y$ by Eqs. (3*a*) and (3*b*), and shows that there is at most one such $x \cap y$.

$$x \geqq x \cap y \qquad y \geqq x \cap y \tag{3a}$$

$$\text{If } x \geqq z \text{ and } y \geqq z\text{, then } x \cap y \geqq z \tag{3b}$$

A lattice is defined as a partially ordered set (poset, for short) in which any two elements x and y have a meet $x \cap y$ and a join $x \cup y$. These binary operations satisfy the four basic identities:

$$\text{(L1)} \qquad x \cap x = x \cup x = x$$

$$\text{(L2)} \qquad x \cap y = y \cap x \text{ and } x \cup y = y \cup x$$

$$\text{(L3)} \qquad x \cap (y \cap z) = (x \cap y) \cap z$$

$$\text{and } x \cup (y \cup z) = (x \cup y) \cup z$$

$$\text{(L4)} \qquad x \cap (x \cup y) = x \cup (x \cap y) = x$$

The operations $\cap$ and $\cup$ are connected with the relation $\leqq$ by the condition that $x \leqq y$, $x \cap y = x$, and $x \cup y = y$ are three equivalent statements. Conversely, if L is an algebraic system with operations $\cap$ and $\cup$ satisfying (L1) to (L4) for all x, y, z, then the preceding condition defines $\leqq$ as a partial ordering of L, with respect to which $\cap$ and $\cup$ have the meanings defined above. This principle was discovered in 1880 by C. S. Peirce.

Kinds of lattices. There are many different kinds of lattices. Thus the real numbers form a lattice if $x \leqq y$ is given its usual meaning. This lattice is simply ordered, in the sense that

$$\text{(P4)} \qquad \text{Given } x \text{ and } y, \text{ either } x \leqq y \text{ or } y \leqq x$$

Any such simply ordered set (or chain) is a lattice, in which $x \cup y$ is simply the larger of x and y, and dually.

Again, the set J of positive integers forms a lattice, if one lets $m \leqq n$ mean "m divides n" (usually denoted $m \mid n$). In this case, $m \cap n = \text{gcd}\ (m,n)$ and $m \cup n = \text{lcm}\ (m,n)$. Still again, one can let Σ consist of all subsets $S, T, \ldots$ of a fixed ensemble I, and let $S \leqq T$ mean that every point in S is in T. Then Σ is a lattice, in which $S \cap T$ is the intersection of S and T, whereas $S \cup T$ is their union. Actually, Σ is a boolean algebra.

In all the preceding lattices, the distributive laws hold:

$$\text{(L5)} \qquad x \cap (y \cup z) = (x \cap y) \cup (x \cap z) \text{ and}$$

$$x \cup (y \cap z) = (x \cup y) \cap (x \cup z), \quad \text{for all } x, y, z$$

Such lattices are called distributive lattices. Any chain is a distributive lattice; so is any boolean algebra. More generally, a ring of sets is defined as a family Φ of subsets of a fixed set I which contains, with any S and T, also their intersections $S \cap T$ and their union $S \cup T$. Then any ring of sets is a distributive lattice.

It is obvious that each of the two identities of (L5) is dual to the other. It is a curious fact that, in a lattice, each also implies the other.

If G is any group, then its subgroups form a lattice, and its normal subgroups also form a lattice, in both cases under inclusion. The normal subgroups satisfy the (self-dual) modular law:

$$\text{(L6)} \qquad \text{If } x \leqq z, \text{ then } x \cup (y \cap z) = (x \cup y) \cap z$$

In general, lattices satisfying (L6) are called modular and every distributive lattice is modular.

Fig. 1. Ordinal number 4.

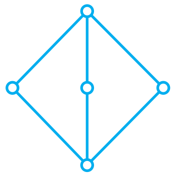


Fig. 2. The projective line over the field Z_2 of integers mod 2.

The lattice of all linear subspaces of the n-dimensional vector space $\mathbf{V}_n(F)$ over any field (or division ring) F is also a modular lattice, usually called the $(n-1)$-dimensional projective geometry $P_{n-1}(F)$ over F. This lattice contains special elements $\mathbf{0}$ (the zero vector) and $I = V_n(F)$ (the whole space), such that

$$\text{(P5)} \qquad \mathbf{0} \leqq x \leqq I \text{ for all } x$$

Such special elements always exist in any lattice whose chains all have finite length, but they need not exist in general; for example, they do not in the simply ordered set of real numbers.

The lattice $P_{n-1}(F)$ is complemented, in the sense that each subspace x has at least one complement x', with the property

$$\text{(L7)} \qquad x \cap x' = \mathbf{0} \text{ and } x \cup x' = I$$

Thus $P_{n-1}(F)$ is a complemented modular lattice. Similarly, it may be verified that the class of boolean algebras is precisely the class of complemented distributive lattices. This principle enables one to consider boolean algebra as a branch of lattice theory.

Lattices L containing few elements can be conveniently visualized by diagrams. In these diagrams, small circles represent elements of L, a being higher than b whenever $a > b$. A segment is then drawn from a to b whenever $a > b$, but no x exists such that $a > x > b$. Any such diagram defines L up to isomophism: $a > b$ if and only if one can travel from a to b along a descending broken line. **Figures 1–6** are typical of such diagrams.

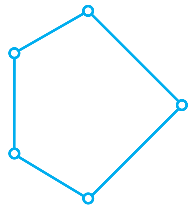


Fig. 3. The simplest nonmodular lattice.

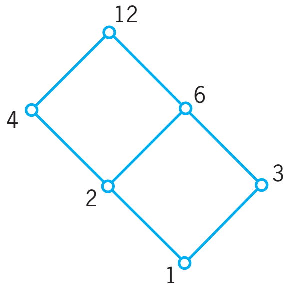


Fig. 4. The lattice of divisors of 12, under divisibility.

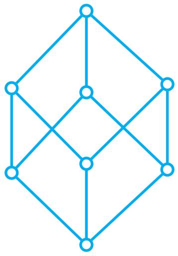


Fig. 5. The boolean algebra of order 8.

Such graphs often give useful information very simply. For example, let a finite lattice be called semimodular if any two elements a and b immediately above (covering) a given element c are also immediately under (covered by) another element $d = a \cup b$. This condition can easily be tested by inspection. Dedekind showed that a finite lattice L was modular if and only if it and its dual were both semimodular. For L to be distributive, the extra condition of containing no subgraph such as that of Fig. 2 is necessary and sufficient.

Applications to algebra and geometry. Lattices, like groups and rings, can be defined as abstract algebras, that is, as systems of elements combined by universally defined operations. These operations may be unary, binary, or ternary. In any such abstract algebra A, define a subset S to be a subalgebra of A when the result of performing any operation of A on elements in S is again in S. Call an equivalence relation $a \equiv b \pmod{\theta}$ on A a congruence relation when, for any n-ary operation f of A, $a_i \equiv b_i \pmod{\theta}$ for $i \equiv 1, \ldots, n$ implies that $f(a_1, \ldots, a_n \equiv f(b_1, \ldots, b_n) \pmod{\theta}$. Then the subalgebras of A form one (complete) lattice, and the congruence relations form another.

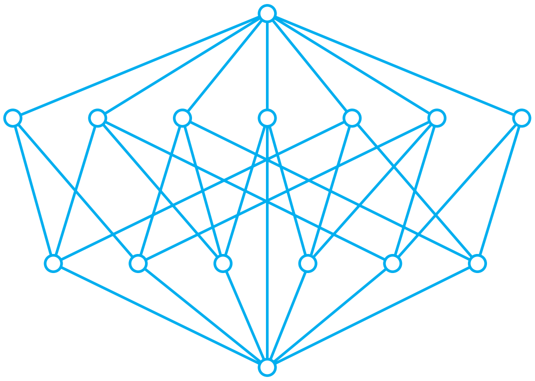


Fig. 6. Lattice of partitions of [*a*,*b*,*c*,*d*].

Such results, which are true of abstract algebras in general, are called theorems of universal algebra. Another such result is the theorem that any algebra can be decomposed into subdirectly irreducible algebras. From this it follows easily that any distributive lattice is isomorphic with a ring of sets.

In groups, rings, and many other algebras, all congruence relations are permutable; it follows that the congruence relations form a modular lattice. This fact, combined with the existence of one-element subalgebras, permits the development of an extensive structure theory that includes a unique factorization theorem under appropriate finiteness conditions. Curiously, lattices themselves satisfy a unique factorization theorem, but for another reason.

It is important to note that the lattice products LM, involved in the statement of the preceding unique factorization theorem, are one of the six basic operations of a general arithmetic of partially ordered sets, which contains the usual cardinal and ordinal arithmetic as a special case. Thus, the most general finite boolean algebra is just $\mathbf{2}^n$, where $\mathbf{2}$ is the ordinal number two and n is a finite cardinal number.

In any lattice with O, a "point" is an element that covers O. In most lattices arising in geometry, every element is a join of a finite number of points; semimodular lattices with this property are called geometric lattices. The lattice of all partitions of a finite set into nonoverlapping subsets is such a geometric lattice. Figure 6 depicts the geometric lattice of all partitions of four objects. Any geometric lattice is complemented; conversely, any complemented modular lattice in which all chains are finite is geometric.

It was mentioned above that the $(n - 1)$-dimensional projective geometry $P_{n-1}(F)$ over a division ring F was always a complemented modular lattice. Many interesting applications of lattice theory to geometry are extensions of this observation. For instance, let P be any abstract projective geometry, defined by incidence relations. Then P is also a complemented modular lattice. Conversely, any finite-dimensional complemented modular lattice is a product of projective geometries and a boolean algebra.

The preceding abstract combinatorial approach to projective geometries led to the construction in 1936, by J. von Neumann, of his continuous-dimensional projective geometries (so-called continuous geometries). O. Frink, in 1946, developed a parallel theory of projective geometries of discretely infinite dimension. By analogy, affine geometries can also be regarded as lattices, and any affine geometry in which all chains are finite is a geometric lattice.

Although important applications of geometric lattices to combinatorial problems have been made by G.-C. Rota and others since 1960, the theory of semimodular lattices in general is quite limited. A negative result is Dilworth's theorem, which states that every finite lattice is isomorphic with a sublattice of a finite semimodular lattice. Similarly, P. M. Whitman has shown that every lattice is isomorphic with a sublattice of the lattice of all partitions of some (infinite) class. *See* ABSTRACT ALGEBRA; COMBINATORIAL THEORY.

Relation to set theory and logic. In set theory, one is frequently concerned with various special families of sets such as closed sets, open sets, measurable sets, and Borel sets. Some of these families form boolean algebras. But the closed sets (and, dually, the open sets) of a topological space X usually form just an uncomplemented distributive lattice $L(X)$. In $L(X)$, the complemented elements (which necessarily form a boolean algebra) are those which disconnect X.

Both the theory of finite sets and the theory of finite distributive lattices are fairly trivial. Each such lattice L has a unique representation as the cardinal power $L = \mathbf{2}^P$ of the two-element boolean algebra $\mathbf{2}$, with a general finite partially ordered set P as exponent. Moreover, any finite-dimensional distributive lattice is finite.

If L is an infinite-dimensional lattice, distributive or not, then one can define several topologies on L, which may or may not be equivalent. Most of these are suggested by a consideration of the special case that L is the rational number system. Adapting Dedekind cuts, as first shown in 1935 by H. M. MacNeille, one can extend any lattice (or partially ordered set) to a complete lattice, in which any subset A of elements x_α has a least upper bound (join) $\vee x_\alpha$ and greatest lower bound $\wedge x_\alpha$. In such a complete lattice, one can define lim sup $\{x_n\}$ and lim inf $\{x_x\}$ for any sequence $x_1, x_2, x_3, \ldots$ of elements. One can define a convergence topology in L by letting $x_n \to x$ mean that lim sup $\{x_n\}$ = lim inf $\{x_n\} = x$. Alternatively, one can define an interval topology by taking the closed intervals $[a, b]$, each consisting of all $x \in L$ with $a \leqq x \leqq b$, as subbasis of closed sets.

The concept of convergence, in lattices and other spaces, can be extended by considering directed sets of indices α, this concept itself being lattice-theoretic. To describe convergence in topological spaces not satisfying countability axioms, such directed sets are essential.

Using the concepts of lattice completeness and lattice topology, one can give interesting characterizations of various lattices associated with topological spaces. Thus, this was done for the lattice of regular open sets by S. Ulam and G. Birkhoff. It was done for the (distributive) lattice of continuous functions by Irving Kaplansky. Finally, it was done for closed (and open) sets by J. C. C. McKinsey and Alfred Tarski, all since 1935.

An interesting analogy between set theory and logic is provided by the equivalence between the algebra of open sets and brouwerian logic, which arose entirely from a consideration of formal logic. Abstractly, both deal with relatively pseudocomplemented distributive lattices.

Another type of logic, directly suggested by quantum theory, was developed in 1936 by G. Birkhoff and von Neumann. In this logic, properties form an orthocomplemented modular lattice. The aristotelian logic of classical mechanics appears as the special, distributive (hence boolean) case of permutable observations. The formal properties of the probability $p[x]$ of an event x are shown to be similar to those of dimension in projective geometry.

Lattice theory also gives perspective on other kinds of formal logic, including modal logic and "strict implication." However, since the deepest problems of formal logic concern the foundations of set theory, these applications will not be discussed here.

Lattice-ordered groups. The first study of lattice-ordered groups was made in 1897 by Dedekind, in analyzing the properties of the rational number system R. Letting $a \leqq b$ signify that $ma = b$ for some integer m (that is, that b is a multiple of a), the nonzero rational numbers form a lattice in which $a \cap b$ and $a \cup b$ again mean the lcm and gcd of a and b. Also, with respect to ordinary multiplication, the nonzero rationals form a (commutative) group. Finally, in R, implications (4*a*) hold; that is, every group translation is a lattice automorphism. Any lattice L which is also a group, and which shares this property, is called a lattice-ordered group, or *l*-group.

$$a \leqq b \text{ implies } ac \leqq bc \tag{4a}$$

$$f(x) \leqq g(x) \text{ implies } f(x) + c(x) \leqq g(x) + c(x) \tag{4b}$$

$$f \geqq g \text{ and } \lambda \geqq 0 \text{ imply } \lambda f \geqq \lambda g \tag{4c}$$

Though many noncommutative *l*-groups exist, it is a striking fact that every complete *l*-group is necessarily commutative.

Lattice-ordered groups arise in function theory, as well as in number theory. If $f \leqq g$ means that $f(x) \leqq g(x)$ for all (or almost all) x, then most function spaces of real functions form lattices. They are also commutative groups under addition. Moreover, implication (4*b*) holds for any $c(x)$; this is simply (4*a*) in additive notation. Hence most (real) function spaces are *l*-groups. In addition, they are vector spaces, in which implication (4*c*) holds. The *l*-groups with these additional properties are vector lattices. Although E. H. Moore and F. Riesz had discussed related ideas earlier, the first systematic analysis of vector lattices as such was made in 1937 by L. V. Kantorovitch.

The application of vector lattice concepts to function theory is still not very far advanced. Using the intrinsic lattice topologies defined earlier, and others related to them, one can avoid the necessity of introducing a distance function, in many function spaces. Thus, the notion of metric boundedness is equivalent to order boundedness for functionals on any Banach lattice, and metric convergence is equivalent to relative uniform star-convergence, which is a purely lattice-theoretic concept.

An additive *l*-group which is also a ring is called a lattice-ordered ring or *l*-ring when its multiplication satisfies the partial analog of implication (4*b*) shown in formula (5). Such rings have been studied systematically only since 1955; a typical theorem about them is the following: An *l*-ring is a product of simply ordered rings if and only if it satisfies implication (6).

$$\text{If} \qquad f \geqq 0 \text{ and } g \geqq 0, \text{ then } fg \geq 0 \tag{5}$$

$$a \cap b = 0 \text{ and } c \geqq 0 \text{ imply } ca \cap b = ac \cap b = 0 \tag{6}$$

Further applications. It is clear that the concepts of vector lattice and of *l*-ring are essential in various physical applications. This was first apparent in connection with the ergodic theorem, as proved in 1931 by Birkhoff and von Neumann, for the deterministic processes of classical mechanics. A generalization of this theorem to stochastic processes, whose natural formulation is based on the concept of a vector lattice, was proved in 1939–1941 by Shizuo Kakutani and Kosaku Yosida.

A second application is to the theory of Reynolds operators, or averaging operators, arising in turbulent fluid motions. The essential connection with the order relation is simply the obvious principle that any average of nonnegative quantities is nonnegative. Using this principle and the theory of *l*-rings, one can decompose (subdirectly) any vector-averaging operator into scalar components.

A third application is to the concept of criticality in nuclear reactor theory. Neutron chain reactions involve the birth (through fission), migration, and death (through absorption) of neutrons. The laws governing the evolution of the statistical distributions of neutrons (as functions of position x, velocity υ, and time t) evidently carry nonnegative distributions into nonnegative distributions. To deduce the mathematical principle that the neutron distribution must satisfy the asymptotic relation $N(x,\upsilon,t) \sim e^{\lambda t}N_0(x,\upsilon)$, λ^{-1} being the reactor "period," it is again most convenient to reformulate the problem in lattice-theoretic language. One can then apply results of Oskar Perron, G. Frobenius, and R. Jentzsch on positive linear operators to prove the desired result. *See* LOGIC; RING THEORY.

Garrett Birkhoff

Bibliography. M. Anderson and T. Feil (eds.), *Lattice-Ordered Groups*: *An Introduction*, 1988; G. Birkhoff, *Lattice Theory*, 3d ed., 1967, reprint 1984; S. D. Cower (ed.), *Universal Algebra and Lattice Theory*, 1985; B. A. Davey and H. A. Priestley, *Introduction to Lattices and Order*, 2d ed., 2001; R. S. Freese and O. C. Garcia (eds.), *Universal Algebra and Lattice Theory*, 1983; G. Gierz et al., *A Compendium of Continuous Lattices*, 1980; G. Gratzer, *General Lattice Theory*, 2d ed., 1998; A. Hoffman, *Continuous Lattices and Their Applications*, 1985; S. Mac Lane and G. Birkhoff, *Algebra*, 3d ed., 1987.

Lattice vibrations

The oscillations of atoms in a solid about their equilibrium positions. In a crystal, these positions form a regular lattice. Because the atoms are bound not to their average positions but to the neighboring atoms, vibrations of neighbors are not independent of each other. In a regular lattice with harmonic forces between atoms, the normal modes of vibrations are lattice waves. These are progressive waves, and at low frequencies they are the elastic waves in the corresponding anisotropic continuum. The spectrum of lattice waves ranges from these low frequencies to frequencies of the order of 10^{13} Hz, and sometimes even higher. The wavelengths at these highest frequencies are of the order of interatomic spacings. *See* CRYSTAL STRUCTURE; VIBRATION; WAVE MOTION.

At room temperature and above, most of the thermal energy resides in the waves of highest frequency. Because of the short wavelength, the motion of neighboring atoms is essentially uncorrelated, so that for many purposes the vibrations can be regarded as those of independently vibrating atoms, each moving about its average position in three dimensions with average vibrational energy of $3kT$, where k is the Boltzmann constant and T the absolute temperature. The wave character of the vibrations is needed, however, to describe heat transport by lattice waves. Also, lattice vibrations interact with free electrons in a conducting solid and give rise to electrical resistance. The temperature variation at low temperatures provides evidence that this interaction is with waves. *See* ELECTRICAL RESISTIVITY.

Lattice waves. In a discrete lattice the elastic equations of a continuum are replaced by a set of linear equations in the displacements $\mathbf{u}(\mathbf{x})$ of atoms at position $\mathbf{x}$ and their accelerations. In a uniform continuum the displacements can be expressed as a superposition of progressive waves, which are defined by their wave vectors, angular frequencies, and polarizations. There are three mutually perpendicular directions of polarization, in general not simply related to the direction of the wave vector. Also, the dependence of the frequency of the wave on the wave vector is linear for a given direction and polarization; the wave frequency is equal to the product of the wave vector and the phase velocity of the wave, which is determined by the elastic moduli and the density of the continuum, and depends on the direction of the wave vector but not its magnitude. *See* PHASE VELOCITY.

In a discrete lattice with perfect periodicity, the normal modes are still progressive waves, and at low frequency they are the same as in the corresponding anisotropic continuum. However, the displacements of atoms are defined only at discrete lattice points, and when the wavelength is no longer large compared to the lattice spacing, the dependence of the wave frequency on the wave vector for a given direction and polarization is no longer linear. This is because the square of the wave frequency can be expressed as a sum of contributions from the force constants of linkages between an atom and its different neighbors, and each contribution contains a factor which varies as the square of the wave vector only when the wave vector is small. There are still three mutually perpendicular polarization directions for every value of the wave vector, but these directions now rotate as the magnitude of the wave vector is changed. This adds complexity to the problem of calculating the frequencies, and numerical methods are required.

The conditions that the lattice be stable and the undisplaced atoms be in positions of equilibrium require that the square of the wave frequency be real and positive for every mode. This implies that for every mode there is a mode of equal frequency and opposite velocity, even in lattices without inversion symmetry. If this were not so, the second law of thermodynamics could be circumvented. *See* THERMODYNAMIC PRINCIPLES.

The dispersion relation, that is, the dependence of the frequency of the wave on the wave vector, determines the density of lattice modes per unit volume and frequency interval.

Zone structure. Like a stroboscopic effect, but in space rather than time, the addition of a reciprocal lattice vector to the wave vector of a particular wave yields the same displacement pattern of the discrete lattice as that of the original wave. Thus, all the lattice waves can be identified with wave vectors in a fundamental zone, bounded by the perpendicular bisector planes of the lowest reciprocal lattice vectors. This fundamental zone contains as many wave-vector values as there are unit cells in the crystal. All real crystals contain at least two atoms per unit cell, and some more. Correspondingly there are, for each wave vector, not only the three polarization modes of the acoustic branches, for which the frequency goes to zero as the wave vector goes to zero (that is, as the wavelength becomes very long), but also the optical branches, which have nonvanishing frequencies everywhere in wave-vector space. When the wave vector equals 0, they describe the relative motion of atoms within a unit cell. In polar crystals they interact strongly with radiation, which accounts for their name. *See* BRILLOUIN ZONE; CRYSTAL ABSORPTION SPECTRA; CRYSTALLOGRAPHY.

Localized vibrations. Dispersion curves of optical modes are frequently flat, that is, the frequency does not vary much over the zone. It thus appears, particularly in Raman spectra, that these modes of vibration have a single frequency, only slightly broadened by interactions with neighboring cells, as if the corresponding vibrations were localized within each cell. In principle, all modes of a harmonic lattice should be wavelike. However, in real crystals there are interactions between modes due to anharmonicities and defects; since these processes can be strong at high frequencies, a model of localized vibrations of atoms relative to each other in a unit cell may sometimes provide a better model. The acoustic modes are always wavelike in crystals. In amorphous solids and highly disordered crystals, they are wavelike at low frequencies but may be localized at the highest frequencies.

Quantization and phonons. A normal mode is like a harmonic oscillator, and according to quantum mechanics its energy is quantized in integral units of hf, where h is the Planck constant. These quanta are termed phonons in progressive lattice or elastic waves. In addition, each mode has a minimum or zero-point energy, $hf/2$; the sum over all normal modes of the zero-point energy is part of the formation energy of the crystal, and not regarded as thermal energy. *See* HARMONIC OSCILLATOR; NONRELATIVISTIC QUANTUM THEORY; PHONON; QUANTUM MECHANICS.

Neutron diffraction. Neutrons are particles, and can also be considered as waves. In a crystal, a neutron can undergo Bragg diffraction, that is, elastic scattering, changing its wave vector by a reciprocal lattice vector, $\mathbf{b}$. In addition, there are weaker inelastic scattering processes, involving absorption or emission of a phonon of frequency f and the wave vector, such that the energy of the neutron increases or decreases by hf and the wave vector of the neutron increases or decreases by the sum of the phonon wave vector and a reciprocal lattice vector. Velocity selection of neutrons scattered in a fixed direction allows their energy and momentum to be determined. Hence, it is possible to obtain the frequency of an interacting phonon of a given wave vector. Through a series of such measurements, the dispersion relation can then be obtained over the entire zone, and the forces between atoms and the density of modes, $g(f)$, can be determined. *See* NEUTRON DIFFRACTION; SLOW NEUTRON SPECTROSCOPY.

Heat transport and wave interactions. Each mode contains thermal energy, and carries a heat current equal to the product of its average energy in thermal equilibrium and the group velocity of the wave. A temperature gradient perturbs the thermal equilibrium, increasing the number of phonons flowing down the temperature gradient. Various interaction processes tend to restore equilibrium, resulting in a steady state with a heat current proportional to the temperature gradient. The constant of proportionality, called the thermal conductivity, involves the contribution to the specific heat per unit volume from waves at various frequencies, the phonon mean free paths or wave attenuation lengths, and the phonon velocities. This relationship applies to isotropic solids and can be generalized for anisotropy. *See* GROUP VELOCITY.

Processes which limit the phonon mean free path are anharmonic interactions, scattering by lattice defects, and scattering by external and internal boundaries. The most important anharmonic interactions are three-phonon processes, in which a phonon is removed from a wave and a phonon is created in each of two other waves, or vice versa. Each such process must satisfy energy conservation as well as a wave-vector interference condition, so that, in the first case, for example, the frequency of the absorbed phonon is the sum of the frequencies of the two created phonons, and the wave vector of the absorbed phonons is the sum of the wave vectors of the two created phonons and a reciprocal lattice vector ($\mathbf{b}$), which may be 0. Three-phonon processes cause the mean free path to decrease with increasing temperature.

Scattering of lattice waves by defects increases with increasing frequency (f); its variation depends on the nature of the defect. Scattering by external and internal boundaries is almost independent of frequency, thus dominating at low frequencies and hence at low temperatures. A study of the thermal conductivity of nonmetallic crystals as function of temperature yields information about the defects present, and about the anharmonic nature of the interatomic forces in the crystal lattice. *See* CRYSTAL DEFECTS; THERMAL CONDUCTION IN SOLIDS.

Paul G. Klemens

Bibliography. N. W. Ashcroft and N. D. Mermin, *Solid State Physics*, 1976; R. Berman, *Thermal Conduction in Solids*, 1976; H. Bilz and W. Kress, *Phonon Dispersion Relations in Insulators*, 1979; G. Grimvall, *Thermophysical Properties of Materials*, 1986, revised 1999; C. Kittel, *Introduction to Solid State Physics*, 8th ed., 2005.

Launch complex

The composite of facilities and support equipment needed to assemble, check out, and launch a rocket-propelled vehicle. The term usually is applied to the facilities and equipment required to launch larger vehicles for which a substantial amount of prelaunch preparation is needed. Small operational rockets may require similar but highly simplified resources on a much smaller scale. For these, the term launcher is usually used. *See* SPACE FLIGHT.

Prelaunch processing. A rocket vehicle consists of one or more stages and a payload. These elements usually are manufactured at different locations and shipped separately to the launch site. The assembly process consists of properly mating these elements in the launch configuration, assuring that all mechanical and electrical interconnections are properly made. Components that have been shipped separately from the main vehicle elements also are installed. These include items requiring special handling or safety precautions, such as batteries and ordnance devices. *See* GUIDED MISSILE; ROCKET STAGING; SPACECRAFT STRUCTURE.

The checkout process consists of detailed testing of all elements of the vehicle to assure they are functioning properly and ready for launch. Checkout of individual components or subsystems may begin prior to or during assembly when easy access for repair or replacement is available. After completion of assembly, the vehicle is given a detailed overall test, usually called combined systems test or integrated systems test, to verify launch readiness.

The launch process includes the final countdown and lift-off. During countdown, propellants are loaded, all vehicle systems are activated and given a final readiness check, ordnance devices are armed, and first-stage motors are ignited.

The term facilities applies to the larger permanent and usually fixed structures on the complex.

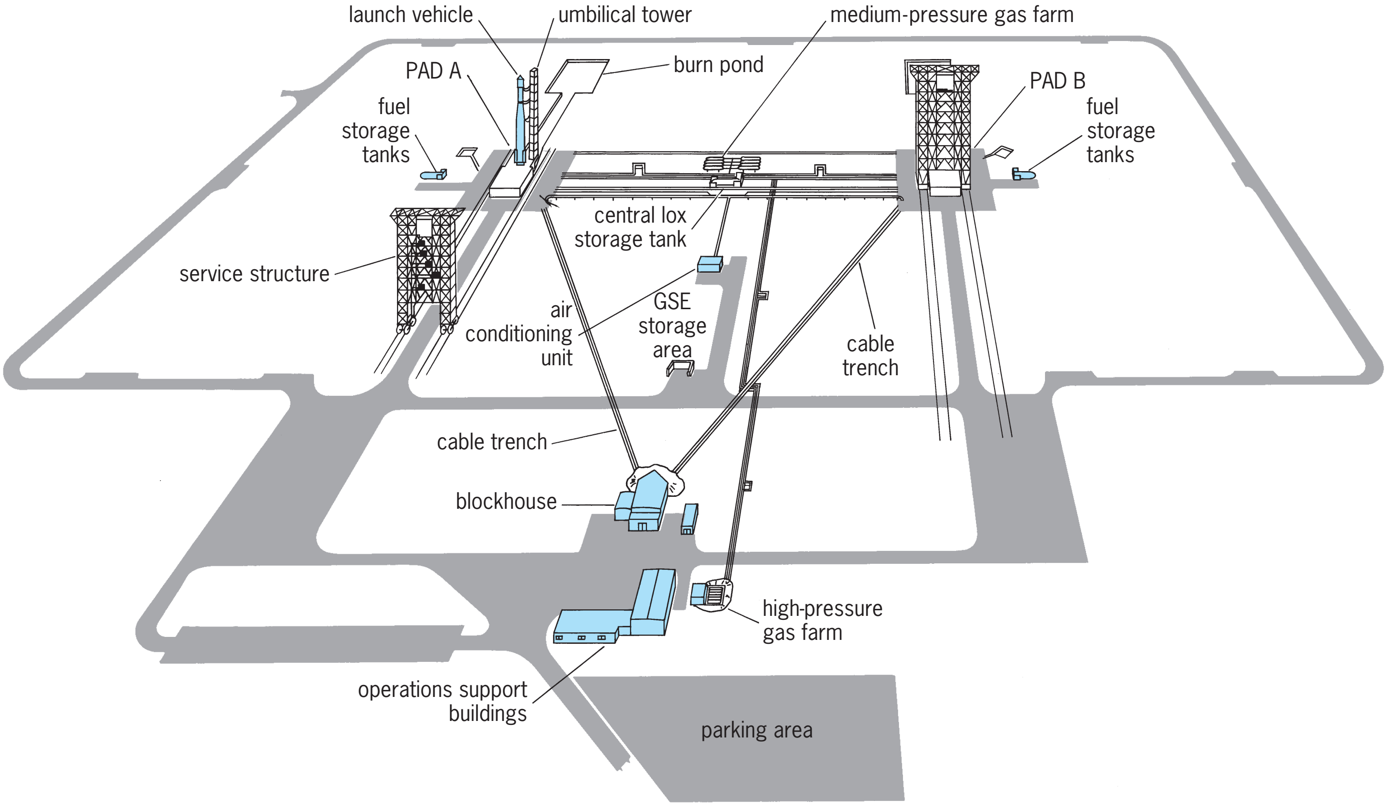


Fig. 1. Overall perspective of a typical launch complex in which one blockhouse serves two launch pads.

The support equipment, often called ground support equipment (GSE), includes all ground equipment necessary to transport, service, test, control, and launch the vehicle. In this grouping are handling equipment, transporters, the service tower and erecting mechanisms, assembly and launch test equipment and consoles, weighing mechanisms, umbilical accessories, control and instrumentation cables, and other external equipment tailored to the missile and required for its preparation and operation.

Launch-complex elements. A typical launch complex (**Fig. 1**) for a vehicle of the Thor-Delta class has two launch pads serviced by a single blockhouse and by common support equipment, providing some economy. The spacing between pads, blockhouse, and other elements of the complex is based on safety considerations. The overall size of a complex therefore depends primarily on the quantities and types of propellants used.

Pad. The launch pad itself is usually a massive concrete structure designed to withstand the heat and pressure of the rocket exhaust. The pad provides physical support for the vehicle prior to launch, but it also has a number of other functions. A propellant-transfer room contains equipment for remote control of propellant loading. An environmental control system supplies conditioned air to the vehicle to maintain normal operating temperatures. A terminal-connection room provides for data and remote-control links between the vehicle and the launch control center or blockhouse.

A flame deflector, located underneath the vehicle, diverts the rocket exhaust in a direction that does not cause harm. A high-pressure, high-volume water supply may be used to cool the flame deflector and otherwise minimize damage to the pad. The water supply also provides fire protection.

Many pads have a hold-down-and-release mechanism that holds the vehicle in place during ignition and releases it when full thrust is built up (**Fig. 2**). The hold-down feature also can be used for test firings of the rocket motors.

The weighing system is a particularly important part of on-stand equipment; it can be used for weighing the missile, determining the amount of propellants loaded, measuring the thrust in static tests, and establishing acceptable thrust performance prior to launching of a missile. Weighing mechanisms are generally load-cell strain gages.

Umbilical connections. During the checkout and countdown phases until just prior to lift-off, the vehicle is serviced from the ground by electrical and mechanical connections. The electrical connections provide electrical power, control signals for remote operation, and data links. The mechanical connections provide for propellant loading, high-pressure gas transfer, and air conditioning. These connections usually are supported by an umbilical tower located on the pad adjacent to the vehicle. Attachment to the vehicle is made by quick-disconnect devices that are pulled away just after rocket motor ignition but prior to lift-off.

Fig. 2. Launch pad with *Soyuz* spacecraft and rocket. Gantry and hold-down mechanisms fold back at proper moments in launch sequence. (*Tass*, *from Sovfoto*)

Service structure. During assembly and checkout, technicians require access to all parts of the vehicle. This is provided by a service structure, or gantry, which can be moved up to the vehicle when needed and pulled away prior to launch. The service

Fig. 3. Minuteman ICBM on the launch pad with its service structure in place. (*Boeing Aircraft*)

Fig. 4. Upper part of a Titan 3 service structure showing the enclosure which provides shelter for the payload. (*TRW Systems*)

structure has movable platforms to provide access to all levels of the vehicle (**Fig. 3**). It may also have means for partially or totally enclosing the vehicle to provide a clean, air-conditioned environment. **Figure 4** shows an environmental enclosure near the top of a Titan 3 service structure intended for sheltering only the payload. **Figure 5** shows the inside of this structure with the hole through which the vehicle will project covered with a safety net.

When the vehicle is assembled on the launch pad, the service structure also provides such necessary mechanical equipment as cranes and hoists, as well as electrical power, gas supplies, and other utilities needed for the process.

Fig. 5. Inside view of the structure of the enclosure shown in Fig. 4. Movable platforms at upper left are temporarily folded away. (*TRW Systems*)

Fueling systems. The launch complex for liquid-fueled vehicles has fuel and oxidizer storage and loading systems. Cryogenic (low-temperature) propellants such as liquid oxygen and liquid hydrogen contribute greatly to the complexity of a launch complex. The cryogens must be stored close enough to the point of use so they will not vaporize in the transfer lines but far enough back from the launch point so that the takeoff blast will not endanger the storage tanks. *See* PROPELLANT; ROCKET PROPULSION.

Extreme cleanliness is necessary in handling liquid oxygen, the most common of oxidizers. Tanks and lines must be thoroughly clean because, under proper conditions, hydrocarbons may combine with an oxidizer to cause a fire explosion. Filters are employed in the system to trap particles which may have been accidentally introduced. Liquid oxygen has a normal temperature of −297°F (−183°C); to minimize losses, the cryogenic tanks are double-jacketed, with a vacuum in the annular space. The entire system is usually made of stainless steel for cleanliness and strength at low temperatures.

Liquid hydrogen, used as a fuel in some vehicles, is even more difficult to handle. It not only has a much lower temperature (−423.2°F or −252.9°C), but hydrogen systems are highly susceptible to leaks. When mixed with air, hydrogen is extremely explosive.

Noncryogenic propellants, such as hydrazine and fuming nitric acid, are used in some vehicles. These are sometimes called storable propellants because they do not boil off at normal ambient temperatures as do cryogens. However, they pose other handling problems because they ignite on contact with each other and are extremely toxic, and fuming nitric acid is highly corrosive.

Fixed fuel storage and pumping areas are ordinarily located on the opposite side of the launch stand from the oxidizer area for safety. For many fueling facilities, the more dangerous and corrosive propellants are handled and transferred by means of specially designed trailers. Even liquid oxygen may be transported in field trailers with their own pumping and pressurization equipment and introduced into the launch vehicle through a loading manifold. Both fuel and oxidizer systems are usually operated remotely from the blockhouse.

A method of dumping propellants is usually provided. Liquid oxygen may be dumped into a concrete evaporation pond, sometimes called a burn pond, on the launch complex. Fuel is rarely dumped on the ground; it is either fed back into its storage tank or piped to a fuel-holding pond where it can be disposed of later. More elaborate closed storage is required for highly toxic or otherwise dangerous propellants.

Control center. The launch operation is controlled from a launch control center. The older control centers, used for uncrewed space missions, are often contained in a blockhouse on the launch complex, a few hundred meters from the launch pad, as in Fig. 1.

Fig. 6. External view of typical blockhouse for uncrewed missions. Massive concrete construction protects workers during launch.

The blockhouse itself is a massive concrete structure (**Fig. 6**) designed to protect personnel from a possible inadvertent impact of the vehicle on the blockhouse with consequent explosion or fire. Thickness of the walls, sometimes 5–7 ft (1.5–2 m) of concrete, is designed to withstand the force of the highest-order explosion that the vehicle propellants could produce. Other precautions, such as baffled underground escape hatches and tunnels and remote air intakes for air conditioning, may be provided.

The launch control center contains both control and monitoring systems. The control systems usually are consoles or panels that enable the operator to remotely operate and control specific functions on the launch pad and in the vehicle. For example, the propellant-transfer console enables the operator to load propellants from the storage tanks to the vehicle. Within the propellant system and the vehicle are sensors that measure flow rates, temperatures, pressures, and other parameters critical to the propellant-loading process. These measurements are transmitted via cable back to the control console, where they are displayed. The operator controls the process and also monitors it to assure satisfactory performance.

The launch control center used for crewed space shuttle missions is located 5 mi (8 km) from the launch pads, far outside the blast area, and therefore does not require a blockhouse. This facility contains four firing rooms, each equipped with consoles which contain controls and displays required for checkout and launch (**Fig. 7**). All four firing rooms are operated by NASA, but two are used to support dedicated Department of Defense missions. The many controls and instruments on the consoles and panels connect to a terminal room on the floor below. Connections are made there to the cables which lead to a similar terminal room at the pad, where in turn connections are made to the vehicle. Closed-circuit television allows direct viewing of many critical operations.

Because many people participate in a launch operation, the communications system that links them together is important to operational success. An operations intercom system may have a dozen or more channels, each channel having a specific assignment. For example, channel 1 may be assigned to stage 1 propulsion, channel 2 to guidance, channel 3 to instrumentation, channel 4 to pad safety, and so on. This scheme of assignment allows each subsystem crew to communicate with each other. The subsystem crew chief also has a link with the system test conductor, for example, with the stage 1 test conductor. The stage 1 test conductor has an additional channel to the chief test conductor, who has overall launch responsibility. The chief test conductor also has direct links to important functions outside the launch complex, such as range control, range safety, and weather forecasting.

Checkout computers. The use of digital computers is important in checkout and launch operations. A typical comprehensive test or launch countdown may require many hundreds of test sequences, each switched on in proper sequence. In addition, many hundreds or even thousands of data measurements

Fig. 7. Typical firing room in the launch control center used for crewed space shuttle missions. (*NASA*)

Fig. 8. Space shuttle Vehicle Assembly Building on Complex 39 at the Kennedy Space Center. The launch control center is at lower left. The space shuttle is on the mobile launch platform in front of the building. (*NASA*)

must be made and evaluated to verify the test results. If this were done manually by human operators, the total time required would be unreasonably long. Computers can control test sequences and data readouts many times faster and with less chance for error. Computers can also promptly evaluate whether a measurement is within a specified tolerance and, if an out-of-tolerance measurement occurs, notify the test conductor. Without the speed and accuracy that computers provide, the checkout and launch of a complex space vehicle, such as a crewed space shuttle, would be virtually impossible. Many rocket vehicle components are deliberately designed for a short life but high performance to achieve weight savings. Such components could wear out faster than they could be checked out in a manual operation. *See* COMPUTER; DIGITAL COMPUTER.

Mobile launcher. Many launch complexes provide for assembly of the rocket vehicle on the launch pad, using the service structure as the assembly facility. Large, complex vehicles such as a space shuttle may require an assembly time of many weeks or months. For them, the launch site is inadequate because the service structure does not usually provide adequate shelter from weather, salt-air corrosion, and other environmental factors; for safety, service shops, stockrooms, engineering offices, and other needed services cannot be located near the launch pad; and the relatively long vehicle preparation time makes the pad unavailable for launchings, thus limiting launch schedules.

To overcome these problems, the mobile-launch concept has been used as the basis for launch-complex design on several programs (Titan 3, Apollo-Saturn 5, and the space shuttle are notable examples). In this procedure, the vehicle is assembled and checked out in specially designed facilities that provide shelter as well as other facilities and support equipment needed for the process. Because highly volatile liquid propellants are not loaded in the assembly building, several vehicles can be safely processed at once, allowing much support equipment to be shared. Usually the vehicle is assembled directly on a transporter which, after the vehicle is assembled and checked out, carries the vehicle to the launch pad and then serves as a launch platform.

The space shuttle Vehicle Assembly Building at the Kennedy Space Center (**Fig. 8**) is 525 ft (160 m) high, 716 ft (213 m) long, and 518 ft (158 m) wide, making it one of the largest (in volume) buildings in the world. A low-bay area is used for overhaul and maintenance of the space shuttle's main engine and serves as a holding area for solid rocket forward assemblies and aft skirts. Solid rocket booster segments and assemblies are processed in adjacent facilities and transferred to the Vehicle Assembly Building for mating with the rest of the complete shuttle vehicle. A high-bay area has four bays, two on each side of the building. Two of the bays are used to store the shuttle's large external fuel tank. The other two bays are used to stack and integrate the complete space shuttle vehicle. Retractable platforms in each bay provide direct access to the vehicle at all work levels. Diagonal to the vehicle assembly building is the launch control center. Two launch pads are located about 4 mi (6 km) distant.

The platform which supports the space shuttle on its mobile launcher (**Fig. 9**) is 160 ft (49 m) long, 135 ft (41 m) wide, and 25 ft (8 m) deep. Two levels within the mobile launcher contain checkout equipment, including a checkout computer, propellant-loading accessories, electrical equipment, and hydraulic servicing units. The mobile launcher serves as the basic support structure for the vehicle during the entire assembly checkout and launch process. The structure, with vehicle aboard, is moved the 4 mi (6 km) from the Vehicle Assembly Building to the launch pad by means of a huge tractorlike device which moves under the mobile launcher, lifts it, and carries it along a specially built roadway.

Space shuttle. Space shuttle operations require two facilities not usually associated with the typical launch complex. The orbiter landing facility is a 3-mi-long (5-km), 300-ft-wide (90-m) runway used for landing operations. The orbiter processing facility

Fig. 9. Fully assembled space shuttle on the mobile launcher leaving the Vehicle Assembly Building. (*NASA*)

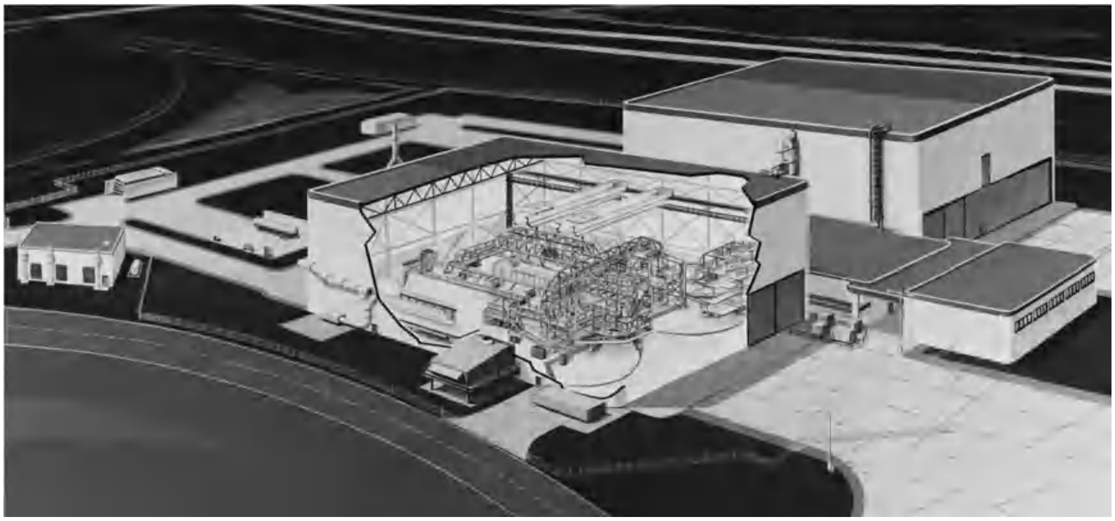


Fig. 10. Orbiter processing facility, used to refurbish orbiter vehicles after return from space and to load and unload heavy cargoes. (*NASA*)

(**Fig. 10**) is used to offload pyrotechnic devices, ordnance items, and toxic fuels, and to refurbish and prepare orbiter vehicles for relaunch after they return from space. The facility is also used to load and unload major payload elements. When vehicle refurbishment and payload operations are completed, the orbiter vehicle is transferred from the processing facility to the assembly building, where it is integrated with the other flight elements, that is, the solid rocket motors and the external fuel tank. *See* SPACE SHUTTLE.

Solid-propellant complex. Solid-propellant rocket vehicles require, in general, a launch complex similar to those just described for liquid-propellant vehicles. The most important difference is the absence of facilities and equipment needed for propellant handling and transfer. This greatly simplifies design of the complex. However, solid-propellant stages are a severe fire or explosion hazard at all times during prelaunch operations. Extreme precautions are taken to eliminate the possibility of accidentally igniting the solid propellants by static electricity, sparks from electrical equipment or machinery, or lightning. Work areas and storage areas are usually separated by revetments.

Operational systems. The launch complex for an operational ballistic missile weapon has essentially the same features as the complexes described above. An operational system is one which is deployed and ready for use by the armed forces. The major differences in the launch complexes are therefore concerned with the operational requirement to maintain a large number of missiles in a launch-ready condition at all times, to have a high probability of successful launches in a short time, and to have a high probability of striking designated targets.

Facilities and equipment for operational systems are designed to have extremely high reliability. Much of the instrumentation used in research-and-development launches is eliminated, and only that which is essential to monitoring launch readiness is retained. A single launch control center may be used for operational control of a large number of missiles, with launch sites for the individual missiles well dispersed around the complex. For weapon systems, solid propellants and storable propellants are much favored over cryogens, which are difficult to store and handle and which cannot be loaded into the vehicle until just prior to launch, greatly increasing the reaction time of the system.

Large weapon-system launch sites may be hardened; that is, they are designed to withstand all but a direct hit from a nuclear weapon in the event of an enemy attack. Hardening usually consists of locating the entire complex underground. The launch pad is a deep hole, called a silo, which contains the vehicle (**Fig. 11**). Tunnels provide limited access for

Fig. 11. Minuteman ICBM emplaced in silo.

umbilicals and at key work levels. As Fig. 11 shows, a technician can be lowered from the surface for additional access, although, on an operational system, access requirements would be minimal.

Perhaps the most unique type of operational launch complex is the Poseidon and Trident-type of submarine. These nuclear-powered vessels can launch 16 Poseidon and 24 Trident intermediate-range ballistic missiles with nuclear warheads while submerged. The relatively short, squat solid-propellant missiles are mounted vertically in injection tubes in a compartment of the submarine. At launch, compressed air ejects the missile to the surface, where the rocket motor is ignited. Complete checkout gear and targeting computers are contained in the submarine. Successful use of the submarine-launched missiles depends on the ability to obtain an accurate geodetic location of the submarine relative to its target. This can be accomplished to a high degree of accuracy anywhere in the world by means of navigation satellites. *See* NAVAL ARMAMENT; SATELLITE NAVIGATION SYSTEMS.

Commercial launch services. With the development of the Space Transportation System, anchored by the space shuttle, NASA began a gradual phaseout of its expendable launch vehicle fleet, consisting of the Atlas-Centaur and the Delta. The last of these vehicles under NASA control were launched in 1989. Under an agreement with the Department of Defense, NASA returned control of Launch Complex 17 (Delta) and Launch Complex 36 (Atlas-Centaur) on Cape Canaveral to the U.S. Air Force. The manufacturers of the Delta and the Atlas-Centaur have agreements with the Air Force for use of these complexes for commercial launch activities involving these vehicles. A similar agreement is in effect between the maker of the Titan vehicle and the Air Force for the use of Complex 40 on Cape Canaveral for commercial launch operations. NASA contracts with the commercial launch organizations for payloads that are not suitable for the space shuttle and that do not require crew-tending during launch, deployment, or orbital operations. Karl E. Kristofferson

Bibliography. *Kennedy Space Center Space Transportation System Ground Operations Plan*, K-STSM-09, Rev. 4, 1979; *Launch Site Accommodations Handbook for STS Payloads*, K-SYSM-14.1, Rev. D, 1989; *Moonport*, SP-4204, 1978; National Aeronautics and Space Administration: *Space Transportation System Facilities and Operations, Kennedy Space Center, Florida*, K-STSM-01, App. A, Rev. A, April 1984; *National Space Transportation System Reference*, vol. 1: *Systems and Facilities*, vol. 2: *Operations*, 1988.

Laurales

An order of flowering plants composed of seven eumagnoliid families of tropical tree species that are important ecologically; some are shrubs. They include in total about 2500 species. They are most closely related to Magnoliales, from which they differ in their partly inferior ovaries and their biaperturate or inaperturate pollen, and then to Winterales and Piperales. Lauraceae (the laurel or cinnamon family) are the best known and largest, but Monimiaceae and its segregates (Atherospermataceae and Siparunaceae) are also important. Nearly all species have aromatic oils, which are important spices, perfumes, and medicines; their flowers are for the most part small and often arranged in distinct whorls, but some such as those of Calycanthaceae are large and much like those of Magnoliales in that parts are arranged spirally and intergrade.

Many species are important as timbers. Cinnamon and camphor come from *Cinnamomum* species, sassafras tea was formerly made from the roots of *Sassafras albida* (now discouraged due to its suspected carcinogenic nature), and avocado comes from *Persea americana*; several genera are cultivated as ornamentals, such as *Calycanthus* (Carolina allspice) and *Chimonanthus* (wintersweet; both Calycanthaceae); and *Laurus* (bay laurel) and *Lindera* (spice bush; both Lauraceae). *See* AVOCADO; CAMPHOR TREE; EUMAGNOLIIDS; MAGNOLIALES; MONOCOTYLEDONS; PIPERALES; SASSAFRAS. Mark Chase

Lava

Molten rock material that is erupted by volcanoes through openings (volcanic vents) in the Earth's surface. Volcanic rock is formed by the cooling and solidification of erupted lava. Beneath the Earth's surface, molten rock material is called magma. All magmas and lavas consist mainly of a liquid, along with much smaller amounts of solid and gaseous matter. The liquid is molten rock that contains some dissolved gases or gas bubbles; the solids are suspended crystals of minerals or incorporated fragments of preexisting rock. Rapid cooling (quenching) of this liquid upon eruption forms a natural volcanic glass, whereas slower cooling allows more minerals to crystallize from the liquid and preexisting minerals to grow in size. The dissolved gases, a large proportion of which are lost on eruption, are mostly water vapor, together with lesser amounts of carbon, sulfur, chlorine, and fluorine gases. With very rare exception, the chemical composition of the liquid part of magmas and lavas is dominated by silicon and oxygen, which form polymers or compounds with other common rock-forming elements, such as aluminum, iron, magnesium, calcium, sodium, potassium, and titanium.

Properties. Viscosity is the principal property which determines the form of erupted lava. It is mainly dependent on chemical composition, temperature, gas content, and the amount of crystals in the magma. Liquid lava with basaltic composition (such as in Hawaii)—relatively low in silicon and aluminum and high in iron, magnesium, and calcium—has higher fluidity (lower viscosity) compared with lava of rhyolitic or dacitic composition (such as at

Mount St. Helens, Washington), with higher abundance of silicon and aluminum but lower amounts of iron, magnesium, and calcium. High temperature and gas content of the liquid lava, combined with low crystal abundance, also contribute to increased lava fluidity. Measured maximum temperatures of basaltic lava (1150–1200°C; 2100–2190°F) are higher than those for andesitic and more silicic lavas (720–850°C; 1330–1560°F). Very fluid basaltic lavas can flow great distances, tens to hundreds of kilometers, from the eruptive vents; in contrast, more silicic lavas travel much shorter distances, forming stubby flows or piling up around the vent to form lava domes. *See* VISCOSITY.

Products. Volcanic products formed by erupted lava vary greatly in size and appearance, depending on volcano type, lava composition, and eruptive style. Most lava products are either lava flows, formed during nonexplosive eruptions by cooling and hardening of flowing lava; or fragmental (pyroclastic) products, formed during explosive eruptions by the shredding apart and ejection into the air of liquid lava.

There are three commonly recognized types of lava flows: pahoehoe, aa, and block lava. Pahoehoe is the Hawaiian name given worldwide to solidified fluid lava whose surface is smooth, gently rolling, and wrinkled to a ropey appearance (see **illus.**); in contrast, aa (also a Hawaiian term) has a very rough, jumbled surface composed of jagged fragments and blocks. Block lava is made up of angular blocks, but bigger and less jagged than those in aa lava.

Fragmental volcanic products differ greatly in size and form, depending on the type of lava ejected and the explosive force of the eruption. All such materials are called pyroclastic. If the explosive force is relatively weak, the lava fragments tend to be larger in size. In contrast, during violently explosive eruptions, the ejected lava is shattered into smaller fragments. However, fragment size in any given explosive eruption can range from room-size blocks to dust-size particles; the largest and heaviest fragments are found closest to the eruptive vent. A great variety of terms have been coined by geologists to describe pyroclastic rocks and deposits, depending on the size, shape, and other characteristics of the fragments. Whether erupted explosively or nonexplosively, basaltic lava generally solidifies to form dark-colored volcanic rocks, whereas more silicic lavas (for example, andesitic, dacitic, and rhyolitic) typically form rocks much lighter in color. *See* ANDESITE; DACITE; IGNEOUS ROCKS; MAGMA; PYROCLASTIC ROCKS; RHYOLITE; TUFF; VISCOSITY; VOLCANIC GLASS; VOLCANO.

Pahoehoe lava (in Kilauea, Hawaii) has a smooth and wavy surface formed by a thin layer of cooler lava at the surface. The surface layer is pushed into folds from the faster-moving, hot fluid lava beneath. (*Photograph by J.D. Griggs, U.S. Geological Survey, June 15, 1989*)

Robert I. Tilling

Bibliography. R. W. Decker and B. B. Decker, *Mountains of Fire*, Cambridge University, New York, 1991; R. W. Decker and B. B. Decker, *Volcanoes*, 3d ed., W. H. Freeman, New York, 1998; P. Francis, *Volcanoes: A Planetary Perspective*, Oxford University Press, Oxford, 1993; G. A. Macdonald, *Volcanoes*: Prentice Hall, Englewood Cliffs, NJ, 1972; H. Sigurdsson et al. (eds.), *Encyclopedia of Volcanoes*, Academic Press, San Diego, 2000.

Lawn and turf grasses

Grasses that were once almost exclusively for pasture and meadow and now furnish ground cover for lawns, sportsfields, industrial parks and roadsides. Acreage of such mowed turf in the United States is reckoned in the tens of millions; in many states the turfgrass industry ranks among the top two or three "agricultural" pursuits. The same species important for grazing are also important for lawns. Grass is well adapted to repeated defoliation (grazing or mowing) because its meristem or growing point is basal, and for the most part only foliage tip growth is sacrificed.

Modern turfgrasses. Although much adventive ("natural") grass is still maintained as lawn, the trend has been toward the breeding and selection of cultivars suited specifically for fine turf, not pasture or meadow. These modern turfgrasses, compared with their "common" antecedents, are typically more decumbent (therefore, retain more green foliage after mowing because of this low stature), are denser by virtue of abundant tillering (hence are more resistant to weed invasion and generally wear better), and are more resistant to pest attack (they are especially screened for disease tolerance). Other features of concern to the turfgrass breeder are seed yields (this correlates a bit negatively with attractiveness), rich color and good looks (fine texture is generally preferred), and ability to spread well (abundant production of rhizomes or stolons). Quickness of seed sprouting, early vigor of the seedling, reluctance to thatch, wear quality, compatibility in blends and mixtures, climatic hardiness, and many other traits are also considered.

The production and marketing of improved turfgrass cultivars is not a static endeavor. In time, pests adapted to once-unaffected cultivar evolve, so new selections are constantly sought. Fortunately, the reservoir of turfgrass breeding lines is not as likely to be exhausted as that of improved agricultural crops, for a diversified germ-plasm bank exists in the many common turfs and the "hedgerows" of suburbia. The parade of new cultivars can be expected to continue, with older varieties dropping out (often because of uneconomical seed yields) and being replaced by still better selections. As an example, representative northern lawn grass cultivars (Lawn Institute Variety Review Board acceptances) are given below:

1. Kentucky bluegrass (*Poa pratensis*); a bit slow from seed, but strong sod, easily cared for.

Adelphi
America
Arboretum
Birka
Bonnieblue
Eclipse
Emmundi
Fylking
Glade
Majestic
Merion
Merit
Monopoly
Nugget
Plush
Ram I
Sydsport
Touchdown
Vantage

2. Perennial ryegrass (*Lolium perenne*): quick starting, as elegant but not so widely adapted as bluegrass; most are polycrosses of several parental bloodlines.

Blazer
Citation
Derby
Diplomat
Fiesta
Manhattan
NK-200
Omega
Pennant
Pennfine
Regal
Yorktown II

3. Fine fescue (*Festuca rubra*, in variety); good for shade, drought and low fertility, but often patchy under warm humidity.

Agram
Banne
Ensylva
Highlight
Koket
Banne
Waldorf

4. Bentgrass (*Agrostis* spp.) and rough bluegrass (*Poa trivialis*): grows in moist shade; mostly for frequently mowed, elegant turfs such as a golf green.

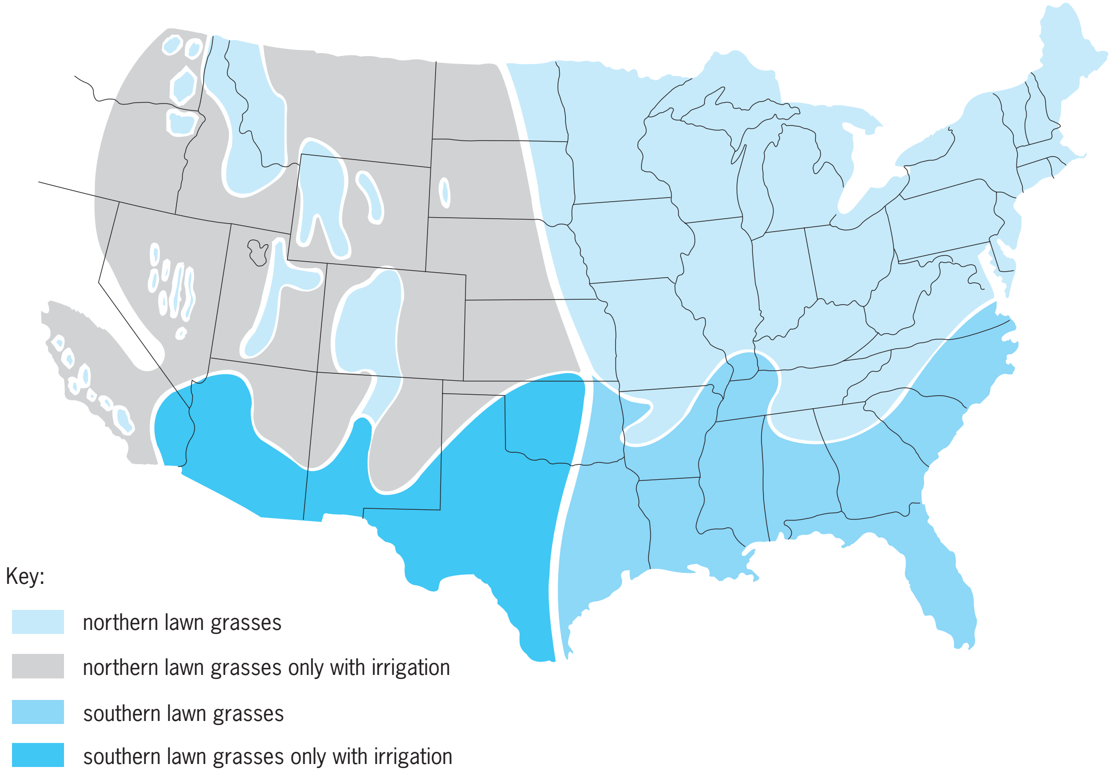
Highland and Exeter colonial bent
Emerald and Prominent creeping bents
Sabre rough bluegrass

5. Tall fescue (*Festuca arundinacea*) "turf-type" polycrosses for difficult border states' "transition zone" climate.

Clemfine, Falcon, and Rebel

Distribution. The **illustration** indicates the general distribution of the major turfgrasses in the United States. The northern species are listed in the table; southern ones are primarily Bermuda grass (*Cynodon*) and manilagrass (*Zoysia*) for the upper South; centipede (*Eremochloa*), St. Augustine (*Stenotaphrum*), and bahia (*Paspalum*) for the deeper South. Tall fescue (*Festuca arundinacea*) is often planted in the "transition zone" where the South meets the North; it is coarse and hardly elegant, but survives well where finer types do not.

As one would suppose, the northern species are at their best during cooler portions of the year. They are especially favored by planting and fertilization in



Major lawn grass zones of the United States.

autumn. Southern species are of a subtropical nature and grow best during the spring and summer.

Turfgrass management. Management practices include mowing, timely fertilization, perhaps irrigation (certainly in arid climates), and probably some measure of pest control.

Mowing will vary with the kind of grass: reel mowers are most suitable for stoloniferous species that are mowed low (bentgrasses and most southern species); the more versatile and less complicated rotaries work better for taller turf. High mowing generally favors turfgrass, since more photosynthetic foliage is retained and the grass is more competitive against weeds. Mowing should be frequent enough so that no more than half of the green foliage is sacrificed at any one clipping.

Fertilization has become costly, and self-reliant cultivars adapted to minimal maintenance are increasingly popular. Still, 2 lb of nitrogen per 1000 ft^2 (1 kg/100 m^2) annually (the experts would recommend at least twice this) should be provided for typical lawns and parceled out during seasons appropriate to growth patterns. Fescues and centipede tolerate low fertility well. Sources of nitrogen that feed out the nutrient gradually are now available in a number of forms, and are to be recommended. In most cases a complete fertilizer (containing phosphorus and potassium as well as nitrogen) would be called for, although for turf the chief and most abundant nutrient should be nitrogen. A soil test will show if phosphorus, potassium, or other cations are adequate. A simple pH test indicates the need for correcting acidity by liming (the humid East) or alkalinity with sulfur (the western plains). *See* FERTILIZER.

Irrigation is essential for survival of traditional turfgrasses in many areas, but prairie species such as buffalograss (*Buchloeë*) and grama (*Bouteloua*) often survive without irrigation. Fortunately, the engineering of lawn irrigation devices has made tremendous strides, and automatic systems are available which apply water precisely according to set schedule. Soils vary in their ability to absorb and retain water, and any irrigation system should be adjusted to their capacities. Sandy soils accept insoak rapidly but may retain only an inch or less in the top foot of soil; they should be watered briefly but frequently. Heavy soils (clays) may accept insoak only slowly, but hold up to 3 in. (8 cm) of water in the pore system of the top foot of soil; they must be watered slowly but less frequently. There is often a tendency to overwater. Actually, letting the lawn dry out occasionally almost to the point of wilting can be therapeutic, helping to forestall hydrophilic weeds (such as nutsedge, *Poa annua*) and disease. *See* IRRIGATION (AGRICULTURE).

Pest control. Weeds are perhaps the most evident pests of lawns. An excellent array of simple-to-use herbicides is available that allows one to selectively control most broadleaf weeds (dicotyledons) in grass, annual grasses in grass, but not perennial grasses in turfgrass (unless all vegetation is killed by a general herbicide and the treated spots reseeded). *See* HERBICIDE.

Insect control is more of a problem, because of the upsetting influence of insecticides on the environment. Insect damage tends to be epidemic, from sod webworms, chinch bugs, soil grubs, and so on. Treatment with a registered insecticide, according to directions, can generally rectify the trouble. *See* INSECTICIDE.

Disease inocula are ever present in a lawn, awaiting proper weather, vulnerable physiology of the grass, appropriate season, lawn stress, and so on. At least a low incidence of diseases such as leaf spot, caused by *Helminthosporium* spp., is almost inevitable in cool spring weather, whereas dollar spot, caused by *Sclerotinia homoeocarpa*, and *Fusarium*-induced diseases are generally manifest in hot weather. Innumerable other diseases can also attack turfgrasses. Most are more intense on heavily fertilized turf, but some (such as dollar spot) are less serious under generous fertilization. Turfgrass diseases are difficult for the amateur to recognize, and even the pathologist is often not sure of causality without undertaking laboratory study and reinoculation of the species.

The concept of ecological "balance" among fungi, with serious disease erupting only when conditions get out of balance, is becoming more accepted. Thus *Fusarium* may be beneficial in the normal lawn microcosm, aiding in decomposition of thatch and holding other diseases in check, although capable of damaging a lawn when a susceptible host and the right conditions coincide. Professional grounds managers who must maintain immaculate turf generally undertake preventive fungicidal sprayings at times of year when disease is likely to erupt. Even then fungi may build resistance to a particular fungicide, so that alternating a conventional contact treatment (which kills spores and mycelia on the surfaces) with the newer systemic fungicides such as benomyl is often advocated. *See* FUNGISTAT AND FUNGICIDE.

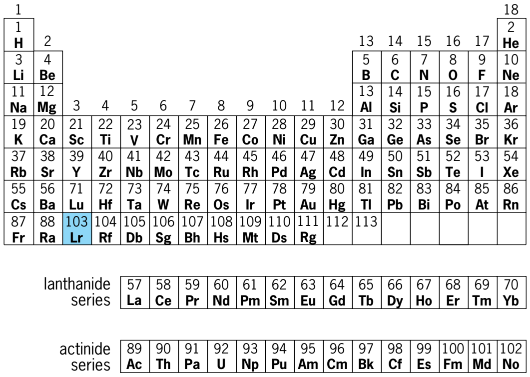
For a homeowner, it is probably too late to stop the inroads of a disease by the time that symptoms are recognized, and in most cases the disease will have to run its course or it may be terminated by changes in weather without any fungicidal treatment. In any event, a homeowner seldom has either the apparatus or the skill to apply a fungicide accurately and frequently enough so that preventive measures work. The most effective measure is to plant disease-resistant lawn grass cultivars. *See* GRASS CROPS; PLANT PATHOLOGY.

Robert W. Schery

Bibliography. J. B. Beard, *Turfgrass Science and Culture*, 1973; W. H. Daniel and R. P. Freeborg, *Turf Manager's Handbook*, 1979; R. W. Schery, *Lawn Keeping*, 1976; A. J. Turgeon, *Turfgrass Management*, 5th ed., 1998.

Lawrencium

A chemical element, symbol Lr, atomic number 103. Lawrencium, named after E. O. Lawrence, is the eleventh transuranium element; it completes the



1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1 **H**																	2 **He**
3 **Li**	4 **Be**											5 **B**	6 **C**	7 **N**	8 **O**	9 **F**	10 **Ne**
11 **Na**	12 **Mg**											13 **Al**	14 **Si**	15 **P**	16 **S**	17 **Cl**	18 **Ar**
19 **K**	20 **Ca**	21 **Sc**	22 **Ti**	23 **V**	24 **Cr**	25 **Mn**	26 **Fe**	27 **Co**	28 **Ni**	29 **Cu**	30 **Zn**	31 **Ga**	32 **Ge**	33 **As**	34 **Se**	35 **Br**	36 **Kr**
37 **Rb**	38 **Sr**	39 **Y**	40 **Zr**	41 **Nb**	42 **Mo**	43 **Tc**	44 **Ru**	45 **Rh**	46 **Pd**	47 **Ag**	48 **Cd**	49 **In**	50 **Sn**	51 **Sb**	52 **Te**	53 **I**	54 **Xe**
55 **Cs**	56 **Ba**	71 **Lu**	72 **Hf**	73 **Ta**	74 **W**	75 **Re**	76 **Os**	77 **Ir**	78 **Pt**	79 **Au**	80 **Hg**	81 **Tl**	82 **Pb**	83 **Bi**	84 **Po**	85 **At**	86 **Rn**
87 **Fr**	88 **Ra**	103 **Lr**	104 **Rf**	105 **Db**	106 **Sg**	107 **Bh**	108 **Hs**	109 **Mt**	110 **Ds**	111 **Rg**	112	113					

lanthanide series	57 **La**	58 **Ce**	59 **Pr**	60 **Nd**	61 **Pm**	62 **Sm**	63 **Eu**	64 **Gd**	65 **Tb**	66 **Dy**	67 **Ho**	68 **Er**	69 **Tm**	70 **Yb**
actinide series	89 **Ac**	90 **Th**	91 **Pa**	92 **U**	93 **Np**	94 **Pu**	95 **Am**	96 **Cm**	97 **Bk**	98 **Cf**	99 **Es**	100 **Fm**	101 **Md**	102 **No**

actinide series of elements. *See* ACTINIDE ELEMENTS; PERIODIC TABLE; TRANSURANIUM ELEMENTS.

The nuclear properties of all the isotopes of lawrencium from mass 255 to mass 260 have been established. ^{260}Lr is an alpha emitter with a half-life of 3 min and consequently is the longest-lived isotope known.

Albert Ghiorso

Bibliography. A. Ghiorso et al., New element: Lawrencium, atomic number 103, *Phys. Rev. Lett.*, 6(9):473–475, 1961; S. Hofmann, *On Beyond Uranium: Journey to the End of the Periodic Table*, 2002; G. T. Seaborg and W. D. Loveland, *The Elements Beyond Uranium*, 1990.

Lawson criterion

A necessary but not sufficient condition for the achievement of a net release of energy from nuclear fusion reactions in a fusion reactor. As originally formulated by J. D. Lawson, this condition simply stated that a minimum requirement for net energy release is that the fusion fuel charge must combust for at least enough time for the recovered fusion energy release to equal the sum of energy invested in heating that charge to fusion temperatures, plus other energy losses occurring during combustion.

The result is usually stated in the form of a minimum value of $n\tau$ that must be achieved for energy break-even, where n is the fusion fuel particle density and τ is the confinement time. Lawson considered bremsstrahlung (x-ray) energy losses in his original definition. For many fusion reactor cases, this loss is small enough to be neglected compared to the heating energy. With this simplifying assumption, the basic equation from which the Lawson criterion is derived is obtained by balancing fusion energy release against heat input to the fuel plasma. Assuming hydrogenic isotopes, deuterium and tritium at densities n_D and n_T respectively, with accompanying electrons at density n_e, all at a maxwellian temperature T, one obtains Eq. (1), where the recovered fusion

$$n_D n_T \langle \sigma v \rangle Q \tau \eta_r \geq \left[\frac{3}{2} kt(n_D + n_T + n_e) \right] \frac{1}{\eta_b} \qquad (1)$$

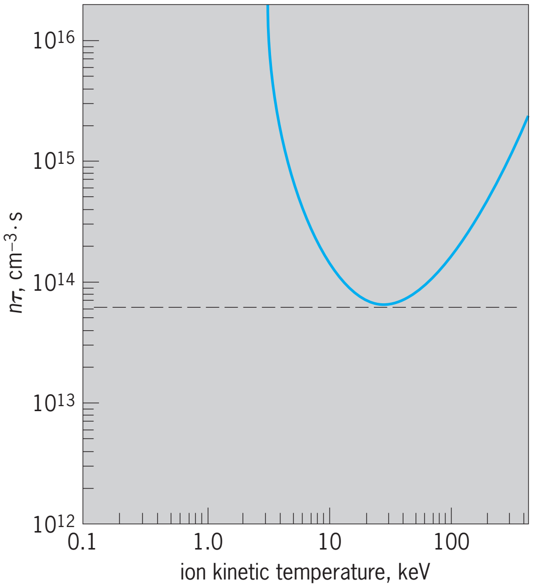
energy release is set equal to or greater than the energy input to heat the fuel. Here $\langle \sigma \upsilon \rangle$ is the product of reaction cross section and relative ion velocity, as averaged over the velocity distribution of the ions, Q is the fusion energy release, η_r is the efficiency of recovery of the fusion energy, η_b is the heating efficiency, and k is the Boltzmann constant.

For a fixed mixture of deuterium and tritium ions, Eq. (1) can be rearranged in the general form of Eq. (2).

$$n\tau \geq F(\eta_r, \eta_b, Q) \left[\frac{T}{\langle \sigma \upsilon \rangle} \right] \qquad (2)$$

For a 50-50 mixture of deuterium and tritium (see **illus.**), the minimum value of $T/\langle \sigma \upsilon \rangle$ occurs at about 25 keV ion kinetic temperature (mean ion energies of about 38 keV). Depending on the assumed efficiencies of the heating and recovery processes, the lower limit values of $n\tau$ range typically between about 10^{14} and 10^{15} cm^{-3} s. These values serve as a handy index of progress toward fusion, although their achievement does not alone guarantee success. Under special circumstances (unequal ion and electron temperatures, unequal deuterium and tritium densities, and nonmaxwellian ion distributions), lower $n\tau$ values may be adequate for nominal break-even.

The discussion up to this point has been oriented mainly to situations in which the fusion reactor may be thought of as a driven system, that is, one in which a continuous input of energy from outside the reaction chamber is required to maintain the reaction. Provided the efficiencies of the external heating and energy recovery systems are high, a driven reactor generally would require the lowest $n\tau$ values to produce net power. An important alternative operating mode for a reactor would be an ignition mode, that is, one in which, once the initial heating of the fuel charge is accomplished, energy directly deposited in the plasma by charged reaction products will thereafter sustain the reaction. For example, in the



Typical plot of minimum value of $n\tau$ necessary for net release of energy versus ion kinetic temperature in a mixture containing equal amounts of deuterium and tritium. (*After J. M. Hollander, ed., Annu. Rev. Energy, 1:213–255, 1976*)

D-T reaction, approximately 20% of the total energy release is imparted to the alpha particle; in a magnetic confinement system, much of the kinetic energy carried by this charged nucleus may be directly deposited in the plasma, thereby heating it. Thus if the confinement time is adequate, the reaction may become self-sustaining without a further input of energy from external sources. Ignition, however, would generally require $n\tau$ products with a higher range of values, and is thus expected to be more difficult to achieve than the driven type of reaction. However, in all cases the Lawson criterion is to be thought of as only a rule of thumb for measuring fusion progress; detailed evaluation of all energy dissipative and energy recovery processes is required to properly evaluate any specific system. *See* NUCLEAR FUSION; NUCLEAR REACTOR; PLASMA (PHYSICS). Richard F. Post

Lawsonite

A metamorphic silicate mineral related chemically and structurally to the epidote group of minerals. Its composition is $Ca(H_2O)Al_2(OH)_2[Si_2O_7]$, and it is a sorosilicate based on the dimeric $[Si_2O_7]^{6-}$ radical, where two (SiO_4) tetrahedra link at a common vertex to form a bow-tie arrangement. The water molecule is bonded directly to the calcium ion and is tightly held. Although the calcium (Ca^{2+}) and the aluminum (Al^{3+}) cations are octahedrally coordinated by anions (and the water molecule), and the silicon cation (Si^{4+}) is tetrahedrally coordinated by the anions, the structure is not limitingly closest-packed.

Lawsonite possesses two perfect cleavages; crystals are orthorhombic prismatic to tabular, and white to pale blue to colorless; specific gravity is 3.1, and hardness is 6.5 on Mohs scale. Lawsonite has symmetry space group Ccmm, $a = 0.890$, $b = 0.576$, $c = 1.333$ nanometers, and four formula units occur in the cell. *See* CRYSTAL STRUCTURE.

Lawsonite possesses an interesting paragenesis (mineral sequence), occurring mainly along continental margins (subduction zones), and it is a good indicator mineral for high-pressure (6–12 kilobars or 600–1200 megapascals) and low-temperature (300–450°C or 570–840°F) assemblages. It is associated with the glaucophane schists (or blueschists) of the Franciscan Formation in California, the Piedmont metamorphics of Italy, and schists in New Zealand, New Caledonia, and Japan. The premetamorphic assemblage is inferred to be largely pelagic and argillaceous (siliceous) sediments derived from basalt weathering and the like. *See* BLUESCHIST; GLAUCOPHANE; SCHIST.

For metamorphism to lawsonite-glaucophane rocks, the precursor fluid phase must be relatively high in water. One reaction curve has a steep pressure slope with lower limit at about 350°C (660°F) and 6 kb (600 MPa) and an upper limit at about 450°C (840°F) and 12 kb (1200 MPa). *See* AMPHIBOLE; EPIDOTE; METAMORPHISM; SILICATE MINERALS.

Paul B. Moore

Layered intrusion

In geology, an igneous rock body of large dimensions, 5–300 mi (8–480 km) across and as much as 23,000 ft (7000 m) thick, within which distinct subhorizontal stratification is apparent and may be continuous over great distances, sometimes more than 60 mi (100 km). Although conspicuous layering may be found in other rocks of syenitic to granitic composition that are richer in silica, the great layered intrusions (complexes) of the world are, in an overall sense, of tholeiitic basaltic composition. (They may be viewed as intrusive analogs to continental flood basalts.) Indeed, their basaltic composition is of paramount significance to their origin. Only basaltic melts, originating in the mantle beneath the Earth's crust, are both voluminous enough to occupy vast magma chambers and fluid enough for mineral layering to develop readily. The relatively low viscosity of basaltic melt is a consequence of its high temperature, 2100–2200°F (1150–1200°C), derived from the mantle source region, and its silica-poor, magnesium- and iron-rich (mafic) composition. *See* BASALT; EARTH; EARTH CRUST; MAGMA.

Cumulus concept. Layered mafic complexes develop upon intrusion of large volumes of basaltic magma (120–24,000 mi^3 or 500–100,000 km^3) into more or less funnel-shaped (smaller intrusions) or dish-shaped (larger complexes) chambers 3–5 mi (5–8 km) beneath the Earth's surface. Although the 1970s and 1980s produced an unusual amount of debate about the development of igneous mineral layering, it is still widely held that such layering is dominantly produced by gravitational settling of early-formed (cumulus) crystals. These crystals begin to grow as the magma cools and, on reaching a critical size, begin to sink because of their greater density relative to that of the hot silicate melt. Although the sequential order of mineral crystallization can vary depending on subtle differences in magma chemistry, a classic sequence of crystallization from basaltic magma is olivine, $(Mg,Fe)SiO_4$; orthopyroxene, $(Mg,Fe)SiO_3$; clinopyroxene, $(Ca,Mg,Fe)SiO_3$; plagioclase $(Ca,Na)(Si,Al)_4O_8$ [**Fig. 1**]. The oxide minerals that are rich in chromium, iron, and titanium, namely chromite, magnetite, and ilmenite, are also common cumulus minerals. Of these silicate and oxide minerals, only plagioclase is light in color. Hence, on various scales, the relative proportion of plagioclase between and within layers is particularly effective in creating the layered aspect of many layered intrusions (for example, **Fig. 2**).

Although layers containing only one cumulus mineral may form under special circumstances, coprecipitation of two or three cumulus minerals—for example, olivine + orthopyroxene; orthopyroxene + plagioclase; or orthopyroxene + clinopyroxene + plagioclase—is more common. Under the influence of gravity and current movements in the cooling, tabular magma chamber, the cumulus minerals accumulate on the ever-rising floor of the chamber. Later recognized as well-formed to rounded grains (**Fig. 3**), these cumulus minerals touch only at occasional

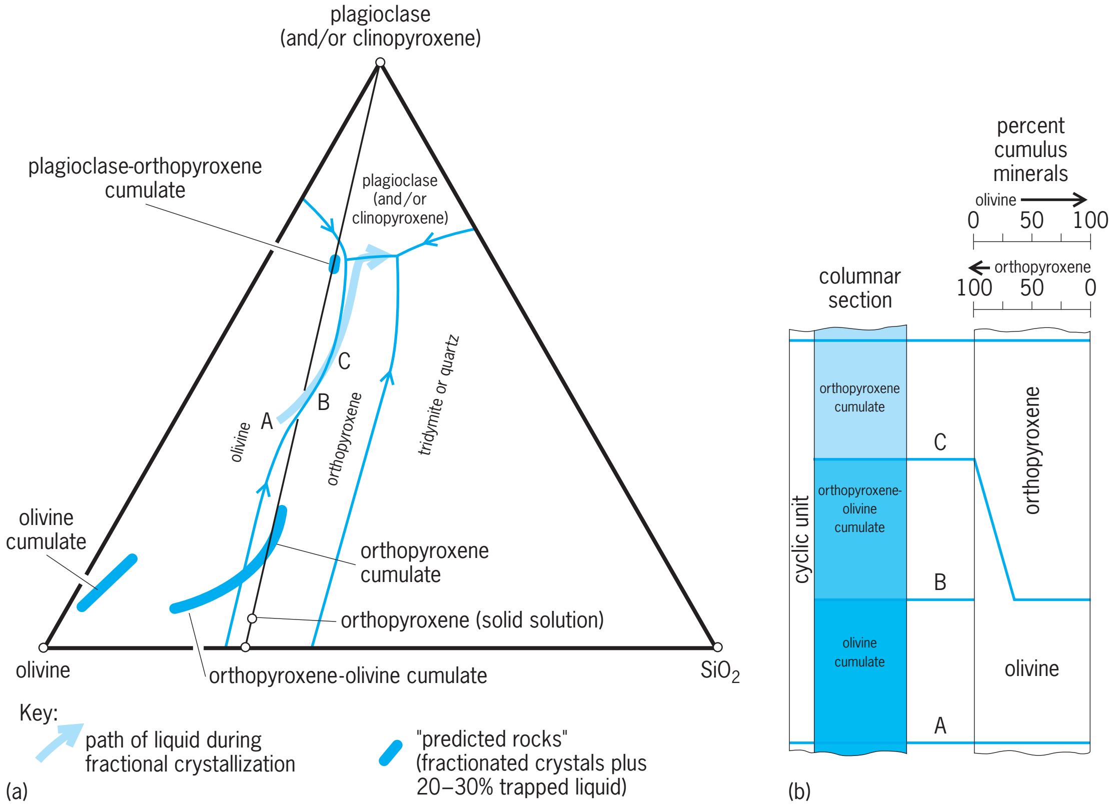


Fig. 1. **Sequence of crystallization from basaltic magma. (*a*) Hypothetical phase diagram showing the derivation of the sequence olivine cumulate, orthopyroxene-olivine cumulate, orthopyroxene cumulate, as observed in cyclic units near the base of many layered mafic complexes. The triangular diagram is divided by cotectic lines (in color) into four fields in which melt compositions can be represented. For melt compositions within each field, only the single mineral for which the field is named will crystallize as the melt cools. Thus, because of withdrawal of the constituents of the crystallizing mineral, the melt changes in composition and migrates to a cotectic line. On reaching a cotectic line, the two minerals for which adjacent fields are named will crystallize together as the magma temperature falls (following the arrowheads on the cotectic lines). Plagioclase-orthopyroxene cumulate consists of plagioclase-clinopyroxene-orthopyroxene or plagioclase-clinopyroxene. (*b*) Hypothetical cyclic unit as derived from a liquid that moved from A to B to C. (*After T. N. Irvine, Crystallization sequences in the Muskox intrusion, in D. J. L. Visser and G. von Gruenewaldt, eds., Symposium on the Bushveld Igneous Complex and Other Layered Intrusions, pp. 441–476, 1969*)**

points and are immersed in residual, interstitial melt.

As cooling proceeds, the interstitial melt typically solidifies as a mosaic of poorly shaped (anhedral) interlocking grains of one or more minerals that form later in the overall sequence of crystallization (Fig. 3). In other cases, particularly when pyroxene and plagioclase are cumulus phases, interstitial melt also may crystallize as overgrowths on cumulus minerals. The solid rock thus formed is known as a cumulate, a term that emphasizes its mode of origin and predominant content of cumulus minerals.

Fig. 2. **A sequence of layers of near-constant thickness and aspect that show a rather abrupt transition from a base rich in pyroxene (pyroxene-plagioclase cumulate) to a top composed nearly entirely of plagioclase (plagioclase cumulate). Ice axe is approximately 3 ft (1 m) long; LaPerouse layered gabbro, Fairweather Range, Alaska.**

Cumulate layering. Lithologic layering within a layered mafic intrusion is typically displayed on a variety of scales. On the broadest scale, such an intrusion may contain ultramafic cumulates rich in olivine and orthopyroxene at its base; mafic pyroxene- and plagioclase-rich cumulates at intermediate levels; and more evolved plagioclase-rich cumulates, or even granitic (granophyric) rocks, near its top. In addition to the relatively conspicuous changes in mineral occurrence and relative abundance, there is generally also a so-called cryptic evolution in mineral chemistry. Early-formed olivine and pyroxene are rich in magnesium (Mg), but become richer in iron (Fe) as successively younger cumulates are deposited; similarly, plagioclase evolves from compositions richer in calcium (Ca) to compositions richer in sodium (Na). Within this broad framework, relatively

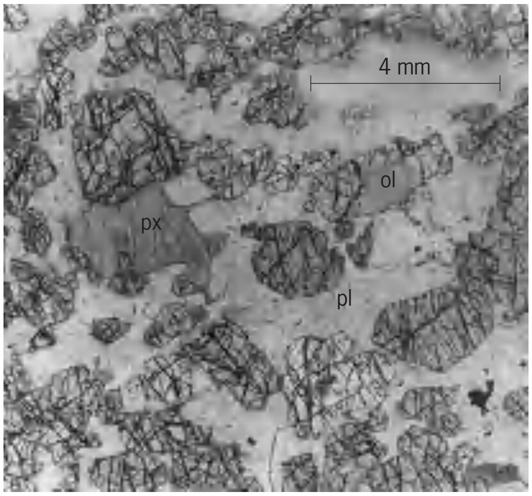


Fig. 3. Photomicrograph of an olivine cumulate from the LaPerouse gabbro. Early-crystallized olivine grains (ol; now fractured and incipiently altered) are enclosed in interstitial plagioclase (pl) and orthopyroxene (px). Mineral proportions suggest that the rock could represent crystallization of a melt with a bulk composition similar to that represented by point A in Fig. 1.

homogeneous cumulate layers range in thickness from 1000 to 2000 ft (300 to 600 m) down to less than 1 in. (2.54 cm). Commonly, layering is rhythmic or repetitious (for example, Fig. 2). The cyclic unit modeled in Fig. 1 may be repeated 15 times in the Ultramafic series of the Stillwater Complex, Montana, where each unit ranges from 80 to 600 ft (24 to 180 m) in stratigraphic thickness.

Mechanics of layering. If a tabular body of basaltic magma were to deposit a cumulate sequence without disturbance, this sequence would not be expected to display many of the observed abrupt changes in cumulus mineral assemblage and chemistry. Thus, commonly observed aberrations in lithologic layering reveal that unusual processes led to their formation. Indeed, many details of these processes have remained speculative.

At least three processes are thought to be commonplace: (1) convective overturn of magma within the chamber; (2) reinjection of new batches of similar magma from the mantle source region; and (3) injection of a chemically distinct, second magma type. In the strictest sense, the first process should produce layered rocks in which mineral proportions and compositions evolve, albeit in a stepwise manner; the second should cause repetition of the initial cumulus assemblage and mineral compositions; while the third should produce such aberrations as monomineralic cumulates, for example, of plagioclase or chromite. One or another of these processes may characterize small layered intrusions, whereas all may have operated to produce the cumulate succession of large complexes. Additional evidence of active current movements within magma chambers are features analogous with sedimentary processes, such as size grading within cumulate layers, cross-bedding, and occasional scour-and-fill structures.

Notable mafic complexes. Geologists have identified several dozen layered mafic complexes, ranging in age from 2700 million years (Stillwater Complex) to 25 million years (LaPerouse, Alaska). By virtue of its enormous size (underlying 25,000 mi^2 or 65,000 km^2) and fabulous mineral wealth, the Bushveld Complex, South Africa, is probably the world's best-known layered intrusion. Quite similar in many respects are the Dufek intrusion of Antarctica (mostly underlying more than 19,000 mi^2 or 49,000 km^2 of glacial ice; **Fig. 4**) and the Stillwater Complex. The Stillwater Complex has been tilted 60° by subsequent tectonic movements, and the 16,000-ft (5000-m) cumulate sequence is now revealed in a 30-mi-long (48-km) cross section. The Great Dyke of Zimbabwe (3.5 by 330 mi or 5.6 by 530 km), the Muskox intrusion of northern Canada, and several complexes in Western Australia are also impressively large. Relatively small complexes such as the Skaergaard intrusion of East Greenland (4.5 by 7 mi or 7 by 11 km), the Rhum intrusion of Scotland, and the Kiglapait intrusion of Labrador (17 by 20 mi or 27 by 32 km) have been studied extensively and have contributed much to the understanding of cumulate processes. The mafic intrusions at Sudbury, Canada, and Duluth, Minnesota, do not qualify as layered complexes in the strictest sense, because of a general absence of well-defined layering.

Economic considerations. Study of layered mafic complexes is of far more than academic interest because many of them host important deposits of chromium (Cr), copper (Cu), nickel (Ni), titanium (Ti), vanadium (V), and the platinum-group elements [platinum (Pt), palladium (Pd), iridium (Ir), osmium (Os), and rhodium (Rh)]. Each of these elements

Fig. 4. Repetitive darker layers of pyroxene cumulates within a thick sequence of relatively monotonous orthopyroxene-plagioclase cumulates, Dufek intrusion, Antartica. Total thickness of cumulates shown is 1750 ft (530 m). (*Art Ford, U.S. Geological Survey*)

has a relatively high initial concentration in mafic magmas, but all must be dramatically concentrated within restricted layers to be recoverable economically. The Bushveld Complex of South Africa has been a dominant source of the world's Cr, Pt, Pd, and V since the 1930s. Significant deposits of Cr in layered complexes are readily recognized because they occur as black, dense, nearly monomineralic cumulate layers of the mineral chromite, $(Cr,Fe)_3O_4$. Such chromite layers may be as thick as 6 ft (2 m) in the Bushveld Complex and have been mapped continuously over distances of tens of miles. Similar, less thick, and continuous layers have been mined from other complexes, including the lower part of the Stillwater Complex, where the chromite is, however, of significantly less value because the proportion of Cr to Fe is substantially lower.

The origin and discovery of platinum-group-element concentrations in layered mafic complexes is more problematic. Indeed, despite intimate knowledge of the Merensky Reef, principal host to the platinoids in the Bushveld Complex, years of intense prospecting preceded discovery of a remarkably similar horizon, the J-M Reef, in the Stillwater Complex. Much of the difficulty experienced in locating the Pt-Pd-enriched horizon in the Stillwater Complex is a reflection of the fact that the Merensky and J-M reefs cannot be readily understood within the framework of the cumulate theory that explains virtually every other feature of these two complexes. It appears that an influx of new magma of undetermined composition, as well as concentration of volatile constituents (such as water and chlorine), trace elements, and platinum-group elements from vast amounts of mafic magma, is essential to formation of a deposit from which Pt and Pd can be recovered profitably. In the Bushveld and Stillwater complexes, the platinum-group elements occur as alloys and in combination with sulfur (S), tellurium (Te), bismuth (Bi), selenium (Se), and other rare elements as minor components of a locally irregular but laterally persistent layer only a few feet thick. *See* IGNEOUS ROCKS.

Gerald K. Czamanske

Bibliography. G. K. Czamanske and M. L. Zientek (eds.), *The Stillwater Complex, Montana: Geology and Guide*, 1985; I. Parsons (ed.), *Origins of Igneous Layering*, 2d ed., 1990; D. J. L. Visser and G. von Gruenewaldt (eds.), *Symposium on the Bushveld Igneous Complex and Other Layered Intrusions*, 1969; L. R. Wager and G. M. Brown, *Layered Igneous Rocks*, 1967.

Layout drawing

A design drawing or graphical statement of the overall form of a component or device, which is usually prepared during the innovative stages of a design. Since it lacks detail and completeness, a layout drawing provides a faithful explanation of the device and its construction only to individuals such as designers and drafters who have been intimately involved in the conceptual stage. In a sense, the layout drawing is a running record of ideas and problems posed as the design evolves. In the layout drawing, for instance, considerations of kinematic design of a mechanical component are explored graphically in incomplete detail, showing only those aspects of the elements and their interrelationships to be considered in the design. In most cases the layout drawing ultimately becomes the primary source of information from which detail drawings and assembly drawings are prepared by other drafters under the guidance of the designer. *See* DRAFTING; ENGINEERING DRAWING.

Robert W. Mann

Lazurite

The chief mineral constituent in the ornamental stone lapis lazuli. It crystallizes in the isometric system, but well-formed crystals, usually dodecahedral, are rare. Most commonly, it is granular or in compact masses. There is imperfect dodecahedral cleavage. The hardness is 5–5.5 on Mohs scale, and the specific gravity is 2.4–2.5. There is vitreous luster and the color is a deep azure, more rarely a greenish-blue. Lazurite is a tectosilicate, the composition of which is expressed by the formula $Na_4Al_3Si_3O_{12}S$, but some S may be replaced by SO_4 or Cl. Lazurite is soluble in HCl with the evolution of hydrogen sulfide.

Lazurite is a feldspathoid but, unlike the other members of that group, is not found in igneous rocks. It occurs exclusively in crystalline limestones as a contact metamorphic mineral. Lapis lazuli is a mixture of lazurite with other silicates and calcite and usually contains disseminated pyrite. Lapis lazuli has long been valued as an ornamental material. Lazurite was formerly used as a blue pigment, ultramarine, in oil painting. Localities of occurrence are in Afghanistan; Lake Baikal, Siberia; Chile; and San Bernardino County, California. *See* FELDSPATHOID; SILICATE MINERALS.

Cornelius S. Hurlbut, Jr.

Le Chatelier's principle

A description of the response of a system in equilibrium to a change in one of the variables determining the equilibrium. Le Chatelier's principle is often stated as the following: For a system in equilibrium, a change in a variable shifts the equilibrium in the direction that opposes the change in that variable. However, this statement is false, since the shift induced by a change in a variable is sometimes in the direction that augments (rather than opposes) the change in the variable. Instead of a single general statement, the response of a system in equilibrium to a change in a variable is best described by a set of particular statements. The following statements are all valid consequences of the laws of thermodynamics. *See* CHEMICAL THERMODYNAMICS.

For any chemical reaction equilibrium or phase equilibrium, an increase in temperature at constant pressure shifts the equilibrium in the direction in which heat is absorbed by the system. For example, the gas-phase reaction $N_2 + 3H_2 \rightarrow 2NH_3$ is exothermic (heat-producing), so an increase in temperature at constant pressure will shift the equilibrium to the left to produce more N_2 and H_2. *See* CHEMICAL EQUILIBRIUM; PHASE EQUILIBRIUM.

For any reaction equilibrium or phase equilibrium, an increase in pressure at constant temperature shifts the equilibrium in the direction in which the volume of the system decreases. For example, the gas-phase reaction $N_2 + 3H_2 \rightleftharpoons 2NH_3$ has fewer moles of gas on the right than on the left, so the products occupy a smaller volume than the reactants, and a pressure increase at constant temperature will shift the equilibrium to the right to produce more NH_3.

For a reaction equilibrium in a dilute solution, addition of a small amount of a solute species that participates in the reaction will shift the equilibrium in the direction that uses up some of the added solute. For an ideal-gas reaction equilibrium, addition at constant temperature and volume of a species that participates in the reaction shifts the equilibrium in the direction that consumes some of the added species.

For an ideal-gas reaction equilibrium, addition at constant temperature and pressure of a species that participates in the reaction might shift the equilibrium to produce more of the added species or might shift the equilibrium to use up some of the added species; the direction of the shift depends on the reaction, on which species is added, and on the initial composition of the equilibrium mixture. For example, for the gas-phase reaction $N_2 + 3H_2 \rightleftharpoons 2NH_3$, addition of a small amount of N_2 at constant temperature and pressure will shift the equilibrium to the left to produce more N_2 if the initial mole fraction of N_2 is greater than 0.5, and will shift the equilibrium to the right to consume some N_2 if the initial N_2 mole fraction is less than 0.5; addition of H_2 or NH_3 at constant temperature and pressure will always shift the equilibrium to consume some of the added species. General conditions determining which way an ideal-gas equilibrium will shift under constant-temperature-and-pressure addition of a species can be derived by using thermodynamics. Ira N. Levine

Bibliography. G. W. Castellan, *Physical Chemistry*, 3d ed., 1983; J. de Heer, Le Chatelier, scientific principle, or "sacred cow," *J. Chem. Educ.*, 35:133, 135–136, 1958; J. de Heer, The principle of Le Chatelier and Braun, *J. Chem. Educ.*, 34:375–380, 1957; I. N. Levine, *Physical Chemistry*, 4th ed., 1994; R. S. Treptow, Le Chatelier's principle, *J. Chem. Educ.*, 57:417–420, 1980.

Leaching

The removal of a soluble fraction, in the form of a solution, from an insoluble, permeable solid with which it is associated. The separation usually involves selective dissolving, with or without diffusion, but in the extreme case of simple washing it consists merely of the displacement (with some mixing) of one interstitial liquid by another with which it is miscible. The soluble constituent may be solid (as the metal leached from ore) or liquid (as the oil leached from soybeans).

Leaching is closely related to solvent extraction, in which a soluble substance is dissolved from one liquid by a second liquid immiscible with the first. Both leaching and solvent extraction are often called extraction. Because of its variety of applications and its importance to several ancient industries, leaching is known by a number of other names; solid-liquid extraction, lixiviation, percolation, infusion, washing, and decantation-settling. The liquid used to leach away the soluble material (the solute) is termed the solvent. The resulting solution is called the extract or sometimes the miscella.

The mechanism of leaching may involve simple physical solution, or dissolution made possible by chemical reaction. The rate of transportation of solvent into the mass to be leached, or of soluble fraction into the solvent, or of extract solution out of the insoluble material, or some combination of these rates may be significant. A membranous resistance may be involved. A chemical reaction rate may also affect the rate of leaching. The general complication of this simple-appearing process results in design by chiefly empirical methods. Whatever the mechanism, however, it is clear that the leaching process is favored by increased surface per unit volume of solids to be leached and by decreased radial distances that must be traversed within the solids, both of which are favored by decreased particle size. Fine solids, on the other hand, cause mechanical operating problems during leaching, slow filtration and drying rates, and possible poor quality of solid product. The basis for an optimum particle size is established by these characteristics.

Leaching processes fall into two principal classes:

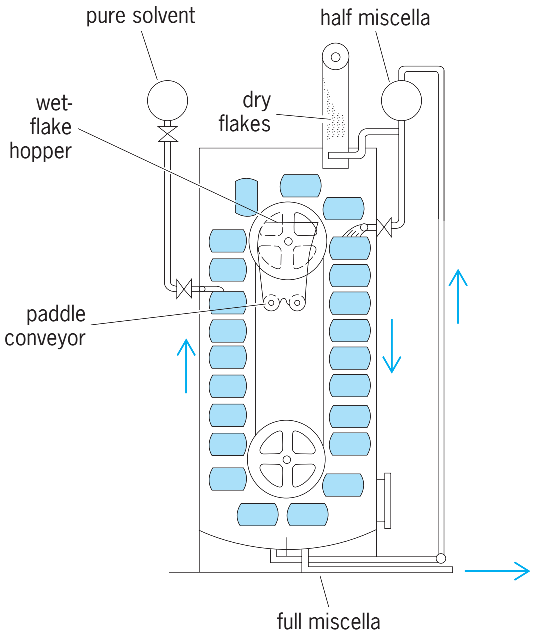


Fig. 1. Bollman extractor. (*After W. L. McCabe and J. C. Smith, Unit Operations of Chemical Engineering, 2d ed., McGraw-Hill, 1967*)

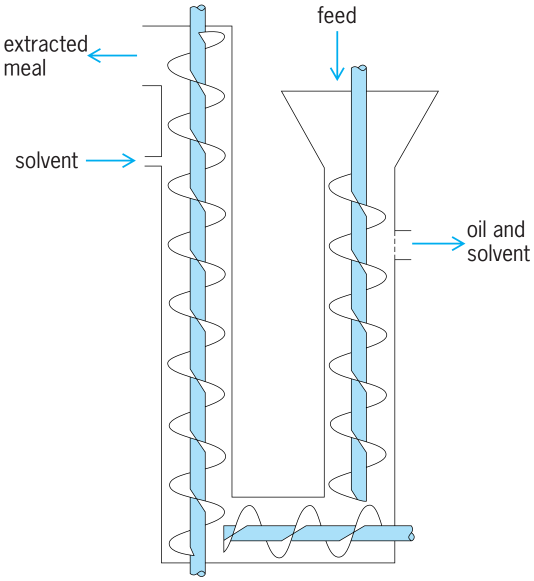


Fig. 2. Hildebrandt screw-conveyor extractor. (*After W. L.McCabe and J. C. Smith*, *Unit Operations of Chemical Engineering*, *2d ed.*, *McGraw-Hill*, *1967*)

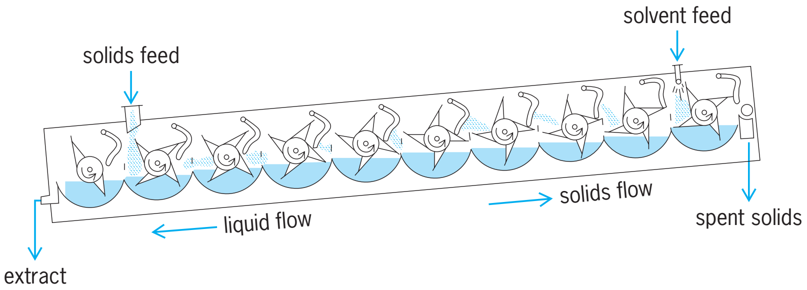


Fig. 3. Kennedy extractor. (*After R. H. Perry and C. H. Chilton*, *eds.*, *Chemical Engineer's Handbook*, *5th ed.*, *McGraw-Hill*, *1973*)

those in which the leaching is accomplished by percolation (seeping of solvent through a bed of solids), and those in which particulate solids are dispersed into the extracting liquid and subsequently separated from it. In either case, the operation may be a batch process or continuous. *See* EXTRACTION; SOLVENT EXTRACTION.

Percolation. In addition to being applied to ores and rock in place and by the simple technique of heap leaching, percolation is carried out in batch tanks and in several designs of continuous extractors.

The batch percolator is a large circular or rectangular tank with a false bottom. The solids to be leached are dumped into the tank to a uniform depth. They are sprayed with solvent until their solute content is reduced to an economic minimum and are then excavated. A simple example is the brewing of coffee in a percolator (repeated extraction) or a drip pot (once-through). Countercurrent flow of the solvent through a series of tanks is common, with fresh solvent entering the tank containing the most nearly exhausted material. Some leach tanks operate under pressure, to contain volatile solvents or increase the percolation rate.

Continuous percolators employ the moving-bed principle, implemented by moving baskets that carry the solids past solvent sprays, belt or screw conveyors that move them through streams or showers of solvent, or rakes that transport them along a solvent-filled trough. In a revolving-basket type like the Rotocel extractor, bottomless compartments move in a circular path over a stationary perforated annular disk. They are successively filled with solids, passed under solvent sprays connected by pumps so as to provide countercurrent flow of the extracting liquid, and emptied through a large opening in the disk. Alternating perforated-bottom extraction baskets may be arranged in a bucket-elevator configuration, as in the Bollman extractor (**Fig. 1**). On the up cycle, the partially extracted solids are percolated by fresh solvent sprayed at the top; on the down cycle, fresh solids are sprayed with the extract from the up cycle. A screw-conveyor extractor like the Hildebrandt extractor (**Fig. 2**) moves solids through a V-shaped line in a direction opposite to the flow of solution. The solids may be conveyed instead by rakes or paddles along a horizontal or inclined trough counter to the direction of solvent flow, as in the Kennedy extractor (**Fig. 3**). In the last two types, the action is predominantly percolation but involves some solids dispersal because of agitation by the conveyors. Horizontal continuous vacuum filters of the belt, tray, or table type sometimes are used as leaching equipment. *See* FILTRATION.

Dispersed-solids leaching. Equipment for leaching fine solids by dispersion and separation, a particularly useful technique for solids which disintegrate during leaching, includes batch tanks and continuous extractors.

Inasmuch as the purpose of the dispersion is usually only to permit exposure of the particles to unsaturated solvent, the agitation in a batch-stirred extractor need not be intense. Air agitation is often used. Examples are Pachuca tanks, large cylinders with conical bottoms and an axial air nozzle or air-lift tube. If mechanical agitation is employed, a slow paddle is sufficient. The Dorr agitator (**Fig. 4**) combines a rake with an air lift. In all cases, the mixture of solids and liquid is stirred until the maximum economical degree of leaching has occurred. The solids

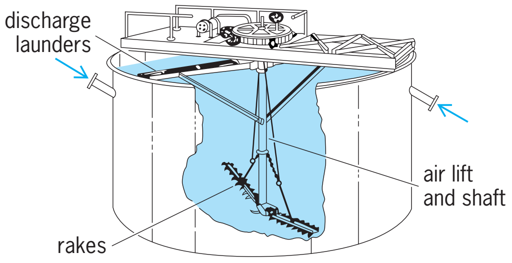


Fig. 4. Dorr agitator for batch washing of precipitates. (*After W. L. Badger and J. T. Banchero*, *Introduction to Chemical Engineering*, *McGraw-Hill*, *1955*)

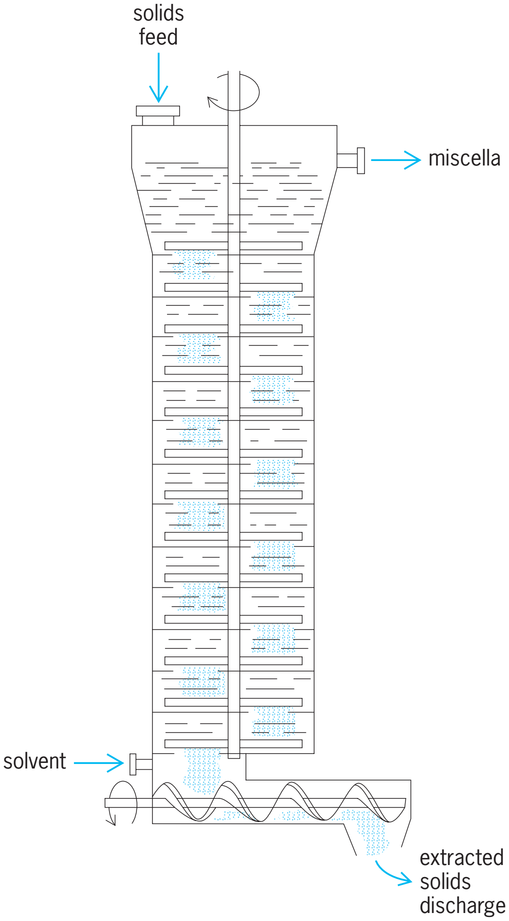


Fig. 5. Bonotto extractor. (*After R. H. Perry and C. H. Chilton, eds., Chemical Engineer's Handbook, 5th ed., McGraw-Hill, 1973*)

are then allowed to settle. The extract is decanted, and the solids, sometimes after successive treatments with fresh solvent, are removed by shoveling or flushing.

Continous dispersed-solids leaching is accomplished in gravity sedimentation tanks or in vertical plate-extractors. An example of the latter is the Bonotto extractor shown in **Fig. 5**. Staggered openings in the plates allow the solids, moved around each plate by a wiping radial blade, to cascade downward from plate to plate through upward-flowing solvent.

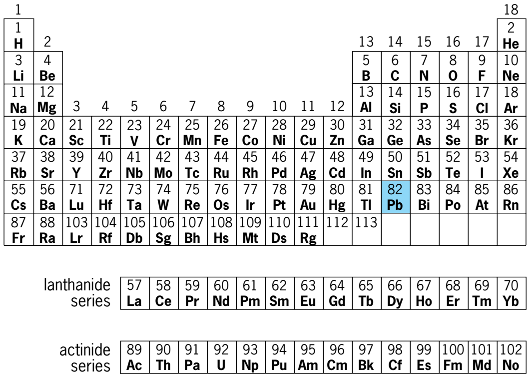
Gravity sedimentation thickeners can serve as effective continuous contacting and separating devices for leaching fine solids. A series of such units properly connected provides true continuous countercurrent washing (known as CCD for continuous countercurrent decantation) of the solids. *See* COUNTERCURRENT TRANSFER OPERATIONS; THICKENING.

Shelby A. Miller

Bibliography. C. J. King, *Separation Processes*, 2d ed., 1980; *Kirk-Othmer Encyclopedia of Chemical Technology*, 4th ed., 1994; W. L. McCabe, J. C. Smith, and P. Harriot, *Unit Operations and Chemical Engineering*, 6th ed., 2000; R. H. Perry and D. Green (eds.), *Chemical Engineers' Handbook*, 7th ed., 1997; J. R. Richardson and D. G. Peacock (eds.), *Chemical Engineering*, 1979; R. E. Treybal, *Mass-Transfer Operations*: *Chemical Engineering*, 3d ed., 1980.

Lead

A chemical element, Pb, atomic number 82 and atomic weight 207.19. Lead is a heavy metal (specific gravity 11.34 at 16°C or 61°F), of bluish color, which tarnishes to dull gray. It is pliable, inelastic, easily fusible, melts at 327.4°C (621.3°F), and boils at 1740°C (3164°F). The normal chemical valences are 2 and 4. It is relatively resistant to attack by sulfuric and hydrochloric acids but dissolves slowly in nitric acid. Lead is amphoteric, forming lead salts of acids as well as metal salts of plumbic acid. Lead forms many salts, oxides, and organometallic compounds. *See* PERIODIC TABLE.



1																	18
1 H	2											13	14	15	16	17	2 He
3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
11 Na	12 Mg	3	4	5	6	7	8	9	10	11	12	13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
55 Cs	56 Ba	71 Lu	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
87 Fr	88 Ra	103 Lr	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112	113					

lanthanide series	57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb
actinide series	89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No

Industrially, the most important lead compounds are the lead oxides and tetraethyllead. Lead forms alloys with many metals and is generally employed in the form of alloys in most applications. Alloys formed with tin, copper, arsenic, antimony, bismuth, cadmium, and sodium are all of industrial importance. *See* LEAD ALLOYS.

Lead compounds are toxic and have resulted in poisoning of workers from misuse and overexposure. However, lead poisoning is presently rare because of the industrial application of modern hygienic and engineering controls. The greatest hazard arises from the inhalation of vapor or dust. In the case of organolead compounds, absorption through the skin may become significant. Some of the symptoms of lead poisoning are headaches, dizziness, and insomnia. In acute cases there is usually stupor, which progresses to coma and terminates in death. The medical control of employees engaged in lead usage involves precise clinical tests of lead levels in blood and urine. With such control and the proper application of engineering control, industrial lead poisoning may be entirely prevented.

Lead rarely occurs in its elemental state. The most common ore is the sulfide, galena. The other minerals of commercial importance are the carbonate, cerussite, and the sulfate, anglesite, which are much more rare. Lead also occurs in various uranium and thorium minerals, arising directly from radioactive decay. Commercial lead ores may contain as little as 3% lead, but a lead content of about 10% is most common. The ores are concentrated to 40% or greater lead content before smelting. *See* LEAD METALLURGY.

The largest single use of lead is for the manufacture of storage batteries. Other important applications are for the manufacture of cable covering, construction, pigments, and ammunition.

Organolead compounds are being developed for applications such as catalysts for polyurethane foams, marine antifouling paint toxicants, biocidal agents against gram-positive bacteria, protection of wood against marine borers and fungal attack, preservatives for cotton against rot and mildew, molluscicidal agents, anthelmintic agents, wear-reducing agents in lubricants, and corrosion inhibitors for steel.

Because of its excellent resistance to corrosion, lead finds extensive use in construction, particularly in the chemical industry. It is resistant to attack by many acids because it forms its own protective oxide coating. Because of this advantageous characteristic, lead is used widely in the manufacture and handling of sulfuric acid.

Lead has long been used as protective shielding for x-ray machines. Because of the expanded applications of atomic energy, radiation-shielding applications of lead have become increasingly important.

Lead sheathing for telephone and television cables continues to be a sizable outlet for lead. The unique ductility of lead makes it particularly suitable for this application because it can be extruded in a continuous sheath around the internal conductors.

The use of lead in pigments has been a major outlet for lead but is decreasing in volume. White lead, $2PbCO_3 \cdot Pb(OH)_2$, is the most extensively used lead pigment. Other lead pigments of importance are basic lead sulfate and lead chromates.

A considerable variety of lead compounds, such as silicates, carbonates, and salts of organic acids, are used as heat and light stabilizers for polyvinyl chloride plastics. Lead silicates are used for the manufacture of glass and ceramic frits, which are useful in introducing lead into glass and ceramic finishes. Lead azide, $Pb(N_3)_2$ is the standard detonator for explosives. Lead arsenates are used in large quantities as insecticides for crop protection. Litharge (lead oxide) is widely employed to improve the magnetic properties of barium ferrite ceramic magnets. Also, a calcined mixture of lead zirconate and lead titanate, known as PZT, is finding increasing markets as a piezoelectric material.

Hymin Shapino; James D. Johnston

Bibliography. American Society for Metals, *Metals Handbook*, 2d ed., 1999; P. W. Atkins et al., *Inorganic Chemistry*, 4th ed., 2006; J. Casas and J. Sordo (eds.), *Lead: Chemistry, Analytical Aspects, Environmental Impact and Health Effects*, 2006; F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; Z. Rappoport (ed.), *The Chemistry of Organic Germanium, Tin and Lead Compounds*, 2003.

Lead alloys

Substances formed by the addition of one or more elements, usually metals, to lead. Lead alloys may exhibit greatly improved mechanical or chemical properties as compared to pure lead. The major alloying additions to lead are antimony and tin. The solubilities of most other elements in lead are small, but even fractional weight percent additions of some of these elements, notably copper and arsenic, can alter properties appreciably.

Cable-sheathing alloys. Lead is used as a sheath over the electrical components to protect power and telephone cable from moisture. Alloys containing 1% antimony are used for telephone cable, and lead-arsenical alloys, containing 0.15% arsenic, 0.1% tin, and 0.1% bismuth, for example, are used for power cable. Aluminum and plastic cable sheathing have replaced lead alloy sheathing in many applications.

Battery-grid alloys. Lead alloy grids are used in the lead-acid storage battery (the type used in automobiles) to support the active material composing the plates. Lead grid alloys contain 6–12% antimony for strength, small amounts of tin to improve castability, and one or more other minor additions to retard dimensional change in service. No lead alloys capable of replacing the lead-antimony alloys in automobile batteries have been developed. An alloy containing 0.03% calcium for use in large stationary batteries has had success.

Chemical-resistant alloys. Lead alloys are used extensively in many applications requiring resistance to water, atmosphere, or chemical corrosion. They are noted for their resistance to attack by sulfuric acid. Alloys most commonly used contain 0.06% copper, or 1–12% antimony, where greater strength is needed. The presence of antimony lowers corrosion resistance to some degree.

Type metals. Type metals contain $2^1/_2$–12% tin and $2^1/_2$–25% antimony. Antimony increases hardness and reduces shrinkage during solidification. Tin improves fluidity and reproduction of detail. Both elements lower the melting temperature of the alloy. Common type metals melt at 460–475°F (238–246°C).

Bearing metals. Lead bearing metals (babbitt metals) contain 10–15% antimony, 5–10% tin, and for some applications, small amounts of arsenic or copper. Tin and antimony combine to form a compound which provides wear resistance. These alloys find frequent application in cast sleeve bearings, and are used extensively in freight-car journal bearings. In some cast bearing bronzes, the lead content may exceed 25%. *See* ANTIFRICTION BEARING.

Solders. A large number of lead-base solder compositions have been developed. Most contain large amounts of tin with selected minor additions to provide specific benefits, such as improved wetting characteristics. *See* SOLDERING.

Free-machining brasses, bronzes, steels. Lead is added in amounts from 1 to 25% to brasses and bronzes to improve machining characteristics. Lead remains as discrete particles in these alloys. It is also added to some construction steel products to increase machinability. Only about 0.1% is needed, but the tonnage involved is so large that this forms an important use for lead. *See* ALLOY; LEAD; LEAD METALLURGY; TIN ALLOYS.

Dean N. Williams

Bibliography. American Society for Testing and Materials, *Nonferrous Metal Products*, vol. 02.04, 1986; W. Hofmann, *Lead and Lead Alloys*: *Properties and Technology*, 1970; T. J. Bertone and J. Neely, *Practical Metallurgy and Materials of Industry*, 5th ed., 1999.

Lead isotopes (geochemistry)

The study of the isotopic composition of stable and radioactive lead in geological and environmental materials to determine their ages or origins. *See* LEAD.

Stable Isotopes

Lead isotope geochemistry provides the principal method for determining the ages of old rocks and the Earth itself, as well as the sources of metals in mineral deposits and the evolution of the mantle.

Geochronology. Lead (Pb) has four stable isotopes of mass 204, 206, 207, and 208. Three are produced by the radioactive decay of uranium (U) and thorium (Th) [reactions (1)–(3), where $t_{1/2}$ is the half-life of

$$^{238}\text{U}\,(t_{1/2} = 4.5 \times 10^9 \text{ years}) \longrightarrow {}^{206}\text{Pb} + 8\alpha + 6\beta \tag{1}$$

$$^{235}\text{U}\,(t_{1/2} = 0.71 \times 10^9 \text{ years}) \longrightarrow {}^{207}\text{Pb} + 7\alpha + 4\beta \tag{2}$$

$$^{232}\text{Th}\,(t_{1/2} = 13.9 \times 10^9 \text{ years}) \longrightarrow {}^{208}\text{Pb} + 6\alpha + 4\beta \tag{3}$$

the isotope and α and β denote alpha and beta particles, respectively]. The lead produced by the decay of uranium and thorium is termed radiogenic. Since ^{204}Pb is not produced by the decay of any naturally occurring radionuclide, it can be used as a monitor of the amount of initial (nonradiogenic) lead in a system. This will include all of the ^{204}Pb and variable amounts of ^{206}Pb, ^{207}Pb, and ^{208}Pb. *See* ALPHA PARTICLES; BETA PARTICLES; RADIOACTIVITY; THORIUM; URANIUM.

Closed systems. It is possible to calculate the isotopic composition of lead at any time *t* in the past by calculating and deducting the amount of radiogenic lead that will have accumulated, provided a mineral or rock represents a closed system. A closed system is one in which there has been no chemical transfer of uranium, thorium, or lead in or out of the mineral or rock since it formed. All calculations for uranium-lead dating should yield the same age; this is a unique and powerful property. The ratio of radiogenic ^{207}Pb to ^{206}Pb is simply a function of age, not the U/Pb ratio. Certain minerals such as zircon, monazite, and uraninite are particularly well suited for dating because of extremely high concentrations of uranium or thorium relative to initial lead. However, the degree to which they behave as closed systems can vary. For samples having concordant U-Pb ages, data lie along a curved line called the concordia, the line defined by the daughter/parent ratios of each isotopic system that have equal ages (**Fig. 1**). There are several uranium-rich minerals that commonly yield concordant ages, the most useful being the rare-earth phosphate monazite, a common accessory mineral in crustal rocks. However, there are many minerals with appreciable radiogenic lead which have discordant ages, indicating they have not been a closed system throughout their history as a discrete phase. Zircon ($ZrSiO_4$), a common accessory mineral in many types of crustal rocks, has been used more than any other phase for U-Pb dating. However, the data are normally discordant. Data for a series of zircons from the Little Belt Mountains, Montana, lie on a well-defined straight line that intersects the concordia at two points (Fig. 1). It has been shown that phases subject to lead loss (or uranium gain) during a period of time that is short compared with the age of the phase yield daughter/parent ratios defining a straight line termed a discordia. The lower intersection of the discordia with the concordia indicates the time of the episodic bulk lead loss, while the upper intersection represents the age of the phase. Discordance in zircons is more pronounced in uranium-rich varieties and is caused by the severe damage to the lattice produced by recoiling alpha particles. *See* MONAZITE; ROCK AGE DETERMINATION; ZIRCON.

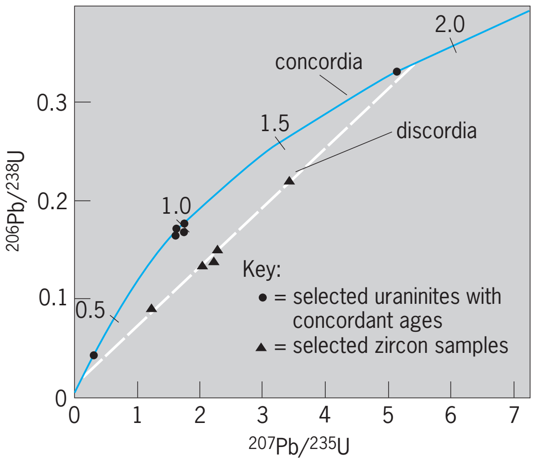


Fig. 1. Systematics of Uranium-lead (U-Pb) dating. For a closed system containing uranium but no primary lead, the ratios of ^{206}Pb/^{238}U and ^{207}Pb/^{235}U will vary with the age of the sample, as shown by the concordia line. Ages indicated by marks along each line are in units of 1 billion years. The data for uraninites are consistent with this. For systems where episodic losses of lead have occurred in the past, values may lie along a discordia, as shown for zircons from the Little Belt Mountains of Montana.

Isochron methods. Even if a rock or mineral contains appreciable initial lead, it may still be dated by using isochron methods. Since the amount of radiogenic lead relative to nonradiogenic lead is a function of the

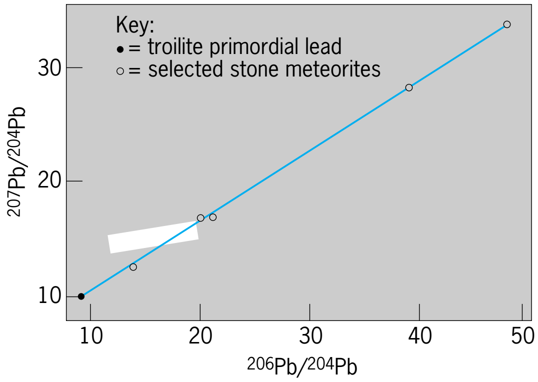


Fig. 2. Plot of ^{207}Pb/^{204}Pb versus ^{206}Pb/^{204}Pb for troilite primordial lead and selected stone meteorites. The slope of the primary isochron (Geochron) for modern lead indicates an age of 4.55 million years for these materials. The white rectangular area illustrates the range of variation in most terrestrial leads and corresponds approximately to the region detailed in Fig. 3.

U/Pb ratio and time, the slope on a plot of ^{206}Pb/^{204}Pb against ^{238}U/^{204}Pb is proportional to age. An isochron is a line on a graph defined by data for rocks of the same age with the same initial lead isotopic composition, the slope of which is proportional to the age. In practice, the ^{238}U/^{204}Pb ratio may well have been disturbed by recent alteration of the rock because uranium is highly mobile in near-surface environments. For this reason it is more common to combine the two uranium decay schemes and plot ^{207}Pb/^{204}Pb against ^{206}Pb/^{204}Pb; the slope of an isochron on this plot is a function of age.

Isochron dating has been used to determine an age of 4.55 billion years for the Earth and the solar system by dating iron and stony meteorites (**Fig. 2**). The position of data along the isochron is a function of the U/Pb ratio. The iron meteorites are particularly important for defining the initial lead isotopic composition of the solar system since they contain negligible uranium. The meteorite isochron is commonly termed the Geochron. *See* EARTH, AGE OF; GEOCHRONOMETRY; METEORITE.

Geochemistry of Earth. By using the position of data for typical continental crustal rocks and samples of basalt that are derived as magmas from the mantle as shown on the Geochron (Fig. 2), the indication is that the silicate Earth has a U/Pb ratio of about 0.1. This is high relative to chondritic meteorites, commonly considered the best representative of primitive unprocessed preplanetary solar system material. A significant fraction of the Earth's total lead inventory could be in the metallic core. Also, lead is extremely volatile and may have been lost at the temperatures that inner solar system objects may have experienced in their accretionary history.

The Earth's mantle has been depleted by repeated melting during its 4.55-billion-year history, and the loss of such melts should leave the mantle with a low U/Pb ratio. However, close inspection reveals that the lead isotopic compositions of most mantle-derived magmas plot to the right of the Geochron (**Fig. 3**), implying a higher U/Pb ratio since the Earth formed. Originally it was thought that this discrepancy was caused by late accretion of the Earth or late core formation, either of which would displace the mantle to the right of the Geochron. However, there is independent isotopic evidence that the Earth did not accrete late, and there are theoretical reasons why the Earth's core almost certainly formed very early. A more likely explanation is that the mantle has been modified throughout its history by the subduction of ocean-floor basalt enriched in uranium and depleted in lead by low-temperature seawater alteration. The basalt lavas of some ocean islands such as St. Helena have especially radiogenic lead, thought to reflect an extreme example of such reenrichment. *See* SUBDUCTION ZONES.

Tracers. Lead isotopes can serve as tracers in the lithosphere, atmosphere, and hydrosphere. Lead isotopes are commonly used to trace the sources of constituents in continental terranes, granites, ore deposits, and pollutants. For example, the class of low-temperature hydrothermal lead-zinc (Pb-Zn) mineralization known as Mississippi Valley type ore deposits have extremely variable ^{206}Pb/^{204}Pb ratios in their galenas, ranging up to 100. These variations reflect the time-integrated U/Pb ratio of the source of the lead, and they can be used to identify specific geological units from which the lead was scavenged. Similarly, some granites such as those of the Isle of Skye in northwest Scotland have very unradiogenic lead, indicating that the magmas were derived by melting portions of the lower continental crust that were depleted in uranium about 3 billion years ago. *See* ORE AND MINERAL DEPOSITS.

The industrialized countries of the world use large tonnages of lead annually, about one-third of which is widely distributed in the air, water, soil, and vegetation of the environment. Isotopic composition of lead in various environmental samples has identified sources and pathways of lead pollution. Most of the lead in the atmosphere originates from the combustion of gasoline containing alkyl lead antiknock compounds. The second-largest emission source of

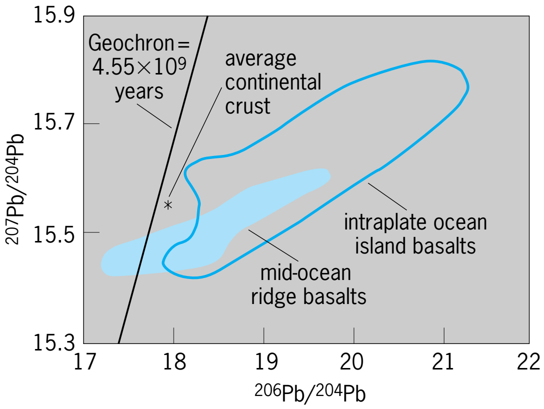


Fig. 3. Lead isotopic compositions of most ocean-floor and ocean-island basalts plot to the right of the Geochron defined by meteorite data (Fig. 2). The composition is the opposite of that predicted from the effects of depletion of the Earth's mantle by partial melting and suggests reenrichment by uranium-enriched subducted ocean floor.

atmospheric lead is coal combustion. Lead aerosols eventually fall to the ground as precipitation or as dust and accumulate in topsoil and in surface water, where they may be incorporated into terrestrial or aquatic life. Lead isotopes have been used to trace contaminant dispersion in the environment. Lead isotope studies, for example, have helped support the contention that high concentrations of lead near roadways are the result of local deposition of large aerosols from automobile exhaust. Similarly, the isotopic composition of lead in natural waters and in sediments has been useful in identifying the extent to which sources are anthropogenic. *See* AIR POLLUTION; WATER POLLUTION.

Radioactive Isotopes

While there are at least 11 known radioactive isotopes of lead, only ^{212}Pb, ^{214}Pb, and especially ^{210}Pb have been of interest geochemically. The usefulness of these isotopes stems from the unique mechanism by which they are separated from parent isotopes in the uranium or thorium decay series.

Atmosphere. Unlike their noble-gas parents, the radioactive lead isotopes as well as other daughter products have a strong affinity for atmospheric aerosols. On formation, the daughter products exist as small positive ions associated with polarized air or water molecules; they form light aggregate particles within periods of tens of seconds. Both ^{212}Pb and ^{214}Pb have been used to study the process of diffusion of ions in gases and the mechanism of attachment of small ions to aerosols. Measurement of the distribution of radon (Rn) daughter product activities (^{212}Pb and ^{214}Pb) with respect to aerosol size has been important in the development of theoretical models of ion-aerosol interactions. The short half-lives of ^{212}Pb and ^{214}Pb also make these isotopes suitable for studies of near-ground atmospheric transport processes.

While the short-lived lead isotopes disappear from the atmosphere primarily by radioactive decay, ^{210}Pb, because of its longer half-life, is removed mainly by precipitation and dry deposition. Its horizontal and vertical distributions are the result of the integrated effects of the distribution and intensity of sources, the large-scale motions of the atmosphere, and the distribution and intensity of removal processes. The inventory of ^{210}Pb in the air is about a thousand times lower than expected, given the amount of its parent ^{222}Rn, a measure of the efficiency with which aerosols are removed from the atmosphere. Numerous measurements of ^{210}Pb as well as other daughter products indicate a tropospheric aerosol residence time of under 1 week. Since the residence time of ^{210}Pb is so short and the oceans are not a significant source, the isotope sometimes can be used to distinguish between air masses originating over the continents and over the oceans. *See* AEROSOL; AIR MASS; ATMOSPHERIC CHEMISTRY; RADON.

Aquatic systems. An important use of ^{210}Pb is as a particle tracer in aquatic systems. Concentrations of ^{210}Pb in surface waves of the oceans generally show the same latitude variations as seen in air and rain. Concentrations in surface waters are roughly 20 times less than expected if there were no removal mechanisms other than radioactive decay. On entering ocean waters, ^{210}Pb is incorporated into microscopic marine organisms, particularly zooplankton, whose remains eventually sink, rapidly conveying ^{210}Pb to underlying waters. Using ^{210}Pb as a tracer has helped explain the mechanisms by which various substances, including pollutants, are removed from the oceans. *See* SEAWATER.

Most of the ^{210}Pb in deep ocean waters is produced from the decay of dissolved ^{226}Ra. The activity of ^{210}Pb is as low as 20% of the activity of radium, indicating the operation of a deep-water scavenging mechanism acting preferentially on ^{210}Pb. Measurements of ^{210}Pb concentrations in deep ocean water indicate a scavenging residence time of around 40 years. The removal process appears to involve horizontal transport of ^{210}Pb to selected areas of intense scavenging by sediments.

One of the most important uses of ^{210}Pb is for dating recent coastal marine and lake sediments. As the isotope is rapidly removed from water to underlying deposits, surface sediments often have a considerable excess of ^{210}Pb. The excess is defined as that present in addition to the amount produced by the decay of radium in the sediments. When the sedimentation rate is constant and the sediments are physically undisturbed, the excess ^{210}Pb decreases exponentially with sediment depth as a result of radioactive decay during burial. The reduction in activity at a given depth, compared with that at the surface, provides a measure of the age of the sediments at that depth. Typically, excess ^{210}Pb can be measured for up to about five half-lives or about 100 years, and it is therefore ideally suited for dating sediments that hold records of human impact on the environment. *See* RADIOISOTOPE (GEOCHEMISTRY); SEDIMENTOLOGY.

Alex N. Halliday; John A. Robbins

Bibliography. R. E. Criss, *Principles of Stable Isotope Distribution*, 1999; A. P. Dickin, *Radiogenic Isotope Geology*, 1997; G. Faure and T. M. Mensing, *Isotopes: Principles and Applications*, 3d ed., 2004.

Lead metallurgy

The extraction of lead from ore, its subsequent purification and processing, and its alloying with other metals to achieve desired properties. Possibly the oldest metal known, lead was particularly attractive to alchemists, predecessors of today's metallurgists and chemists, as a substance which they believed could be transformed into gold.

There are several ores of lead; the most important is galena, PbS. Other ores are cerussite, $PbCO_3$, and anglesite, $PbSO_4$. Lead ores are often accompanied by zinc ores, and galena usually contains profitable quantities of silver and gold. *See* CERUSSITE.

Specifications and grades. There are many types and several specifications for lead. In the United States, these include ASTM B29 and Federal Specification QQ-L-171; in Germany, DIN Standard 1719;

Chemical requirements for various grades of lead

	United States ASTM Pig Lead B29				Germany Specification DIN 1719				United Kingdom Specification BS334	
Element	Corroding	Chemical	Acid-copper	Common desilverized	Pure Pb 99.99	Pure Pb 99.985	Smelter lead	Pure copper lead	Chemical type A	Chemical type B
Ag, max %	0.0015	0.020	0.002	0.002	0.001	0.001	0.001	0.0025	0.002	NS*
Ag, min %	—	0.002	—	—	—	—	—	—	—	NS
Cu, max %	0.0015	0.080	0.080	0.0025	0.001	0.001	0.001	0.08	0.003	NS
Cu, min %	—	0.040	0.040	—	—	—	—	0.04	—	NS
Ag + Cu, max %	0.0025	—	—	—	—	—	—	—	—	NS
As, max %	—	—	—	—	0.001	0.001	0.001	0.001	Trace	NS
Sb, max %	—	—	—	—	0.001	0.002	0.002	0.002	0.002	NS
Sn, max %	—	—	—	—	0.001	0.001	0.001	0.001	Trace	NS
As + Sb + Sn, max %	0.002	0.002	0.002	0.005	—	—	—	—	—	NS
Zn, max %	0.001	0.001	0.001	0.002	0.001	0.001	0.001	0.001	0.002	NS
Fe, max %	0.002	0.002	0.002	0.002	0.001	0.001	0.001	0.001	0.003	NS
Bi, max %	0.050	0.005	0.025	0.150	0.005	0.01	0.05	0.01	0.005	0.005
Cd, max %	—	—	—	—	—	—	—	—	Trace	NS
Co, max %	—	—	—	—	—	—	—	—	0.001	NS
Ni, max %	—	—	—	—	—	—	—	—	0.001	NS
Pb, min %	99.94	99.94	99.90	99.85	99.99	99.985	99.94	99.90	99.99	NS

*NS = not specified; open to negotiation between supplier and customer.

and in the United Kingdom, BS 334. The lead refined to meet these specifications is usually primary lead, that is, lead which is extracted and refined from ore. Secondary lead, recycled mostly from old storage batteries, can be refined to meet the same specifications as the primary metal, but the remelting, drossing, and recasting involved may have a negative impact on composition.

In the United States most lead ore is smelted and refined to a minimum purity of 99.85%. At and above this level of purity, four different grades of lead are recognized by ASTM B29: corroding, chemical, acid-copper, and common desilverized lead. The major differences among these grades are the allowable concentrations of copper, silver, and bismuth. Even trace amounts of these elements can have a significant effect on the properties or cost of the lead and justify having the four grades. Other elements whose concentrations are generally kept low include antimony, arsenic, tin, zinc, and iron. The chemical composition requirements for the four ASTM grades of lead are given in the **table**. The **illustration** shows an abridged flowsheet of lead metallurgy.

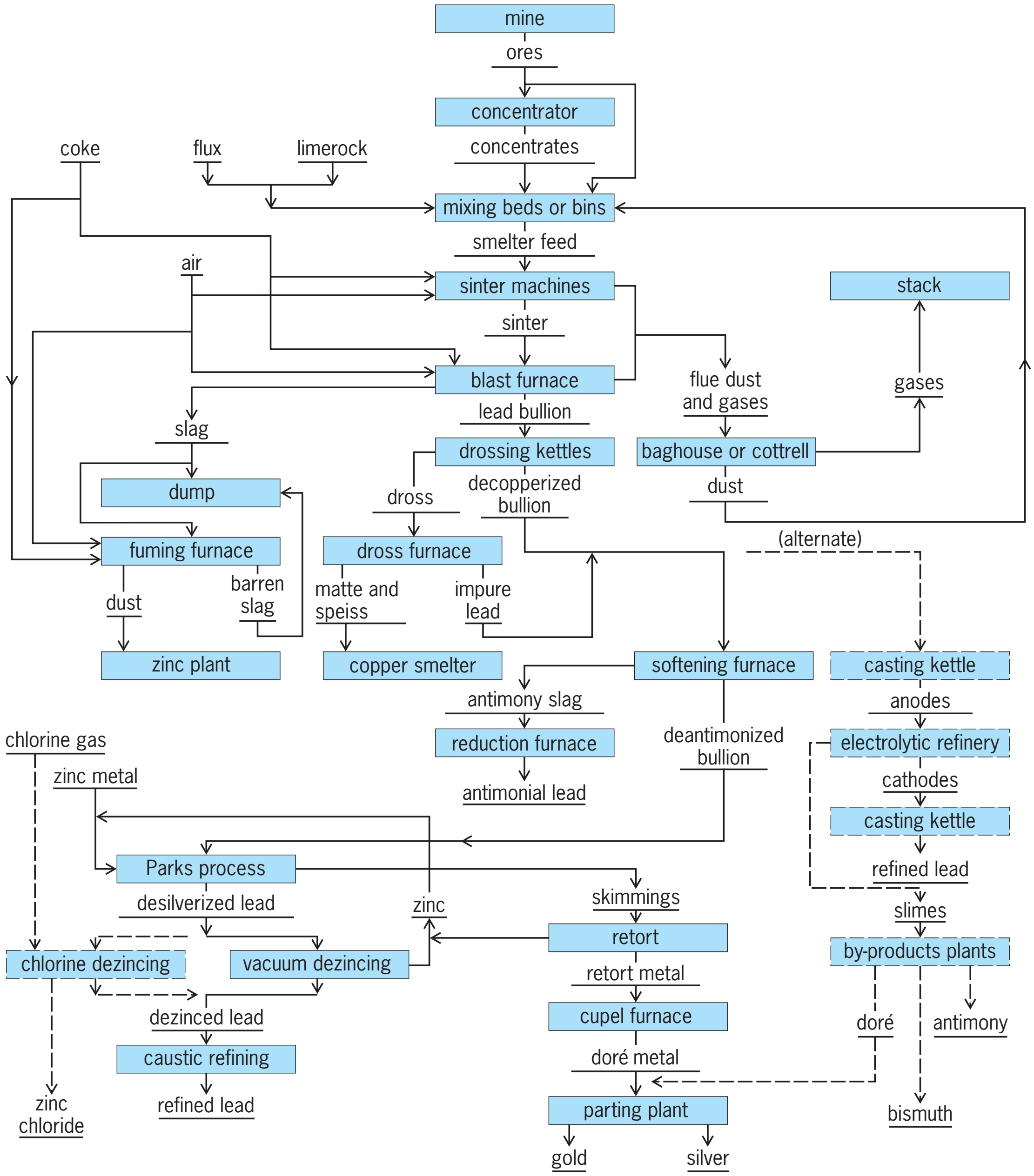
Concentrating. The first step in the beneficiation of ores to raise the lead content and to separate the lead from the zinc and iron minerals is concentrating. The standard processing begins with crushing the raw ore, followed by wet grinding, which reduces the product to a particle size of 75% minus 325 mesh. The resulting slurry is conditioned with certain reagents to establish a proper alkalinity, and by further mixing with flotation chemicals that collect the lead minerals in a froth, which is thickened and filtrated. Reagents include sodium carbonate, lime, copper sulfate, pine oil, cresylic acid, xanthate, and sodium cyanide, among others, in amounts ranging from 0.05 to 5.0 lb per ton (0.25 to 2.5 kg per metric ton) of ore. *See* FLOTATION.

During the lead flotation process, zinc minerals, iron compounds, and earth components of the ore are depressed instead of floated by this treatment, and are recovered after further separations. Copper, silver, and gold, if present, normally remain with the lead and are removed and recovered in the refinery. However, copper, if present in sufficient quantity, may be separated from the lead-copper concentrates by special flotation procedures.

The lead concentrate produced in this first processing step has a lead metal content of about 70%.

Smelting. Before smelting, lead concentrates are frequently blended with high-grade raw ores or returned intermediates (such as flue dusts, limerock, and so forth) which are drawn from proportioning bins. These materials are pelletized so that a homogeneous and carefully sized smelter feed is provided. The feed is then sintered to eliminate most of the sulfur and to agglomerate the particles into relatively large, hard lumps that will not be blown out of a blast furnace.

Sinter machines consist of a chain of moving pallets on which the porous feed is ignited. The ore pellets are subjected to an air blast (downdraft or updraft) which burns the sulfur, creating sulfur dioxide, and at the same time oxidizes the metallic elements. The sinter product, with about 9% of its weight of carbon in the form of coke, is charged into the top of a blast furnace. This is a simple vertical shaft that could be as high as 23.3 ft (7.11 m) and with a 5.5-ft (1.68-m) diameter. The coke supplies the fuel for melting the charge, and also the reducing gas which reacts with the lead oxide to form metallic lead. As the charge descends in the furnace, the molten metal flows to the bottom, from where it is



An abridged flowsheet of lead metallurgy.

withdrawn for further treatment. The remainder of the charge forms slag that floats on the lead and is removed from the furnace at a higher level. An improvement is the continuous tapping of slag, which eliminates the high labor requirement for the intermittent operations of the past.

The lead bullion contains the silver, gold, copper, bismuth, antimony, arsenic, tin, and other minor metals in the ore; the slag carries the zinc, iron, silica, lime, and other gangue; the dust contains a variety of elements, including cadmium and indium, plus much lead and zinc. A typical blast furnace takes about 750 tons (680 metric tons) of sinter per day and produces some 250 tons (230 metric tons) of lead bullion. *See* SINTERING.

Refining. The impure bullion is cooled in kettles to about 660°F (350°C), causing a dross to form which carries almost all of the copper as well as major amounts of lead. Most of the copper segregates in the kettle at low temperatures because the solubility of copper in lead, just above the latter's melting point, is very low. After the dross, which has a high copper content, is skimmed off, the rest of the copper is removed by stirring sulfur into the bath. At this stage, the work lead has been decoppered down to 0.01%. If significant quantities of tin are present, the lead bullion may be reheated to 1100°F (600°C), with the introduction of air, and a second dross containing the tin is removed. The lead bullion is then sent on to the refinery.

Slag from the blast furnace is frequently treated by fuming out the zinc and remaining lead. The barren slag is discarded. The dusts from the sinter plant and blast furnace are collected in a baghouse or precipitator, and returned to the mixing bins or proportioning bins.

Decopperized lead bullion, containing significant amounts of silver, gold, and other materials, is refined again by one of two principal processes. Most of the world's output is produced by means of pyrometallurgical techniques, while the remainder—less than 20%—is electrolytically refined. The latter process is used only when the bismuth content of the lead bullion is relatively high.

The first step in the pyrometallurgical operation is to soften the metal by removing the arsenic, antimony, and tin. This can be done with air oxidation of these elements in a small reverberatory furnace at a temperature of 1300–1380°F (700–750°C). The slag, which is normally high in antimony and lead content, is reduced with coke, and the resulting antimonial lead alloy is marketed. Another widely used method of removing these impurities is the Harris process. In this operation, liquid lead bullion is sprayed through molten caustic soda and molten sodium nitrate. The latter reagent oxidizes the arsenic, antimony, and tin, which are converted into sodium salts and skimmed from the bath. The three elements can be recovered from these compounds by wet chemical methods.

After the refining process has been carried to this point, the silver and gold are separated, using the Parkes zinc-desilverizing process. This method is based on the principle that when molten lead is saturated with an excess of zinc the precious metals become insoluble in lead. Intermetallic compounds with zinc are formed, and these float to the surface as a solid crust or dross which can be skimmed off. After two or three such treatments, the lead is free of silver and gold and proceeds to the next operation. The desilverized lead is essentially a eutectic melt of lead and zinc, and the latter is now usually extracted by vacuum dezincing, in which the main part of the zinc is evaporated under vacuum at 1100°F (600°C) and recovered in metallic form. Some plants dezinc the lead by passing chlorine through the metal, forming zinc chloride. The residue is essentially an impure alloy of silver and gold. After an oxidation treatment in a cupel furnace, the now partly refined alloy, termed doré metal, is electrolytically parted into refined silver and gold.

Grades of lead with a high bismuth content are unsuitable for some purposes, and so removal of the bismuth is economical for contents above 0.1%. Minor quantities of bismuth are likely when lead is refined by pyrometallurgical means. The bismuth may be eliminated by the Betterton-Kroll process, which involves treatment of the melt with small amounts of calcium and magnesium. These combine with the bismuth and rise to the surface as a dross which is readily skimmed. There are other methods for reducing the bismuth content to as little as 0.001% by precipitation with sodium. Removal of the bismuth is also possible by electrolytic refining of lead, but economically so only when the bismuth content exceeds 0.5%.

Lead produced by the processes described, or by others, can have a purity of 99.90 to 99.99%. At this level the grades of lead differ only by their bismuth content. Smelting techniques make it possible to produce commercial lead of any desired degree of purity, independent of the nature of the raw material. Lead with so-called six nines purity (99.9999) can be provided for special situations or scientific research. *See* ELECTROLYSIS; PYROMETALLURGY, NONFERROUS.

Alloys. If a user requires certain alloys to achieve desired properties, the elements are added later. In general, addition of small amounts of other metals increases the hardness and tensile strength of lead and decreases its ductility, malleability, and density. Often the tendency for lead to creep, its melting point, and the grain size are also reduced. Subsequent additions of the same metal do not usually provide a linear improvement in the mechanical properties. After a certain amount of alloying metal has been added, further additions often create alloys which are weaker and more brittle.

Another factor which depends on alloy concentrations is the tendency of certain alloys to improve in time due to age hardening. However, alloys that age-harden are also susceptible to overaging, which turns strong alloys into the original soft leads, with the alloying metal ending up in segregated concentrations. Heating can accelerate both age hardening and overaging.

The percentage of alloying metal which will provide a specific amount of improvement in mechanical properties depends in part upon the difference between the atomic radius of lead and that of the alloying metal. It also depends on whether the alloy is heat-treated and whether it is formed by casting, rolling, or extruding. Usually the greater the difference in atomic radii between the lead and the second metal, the greater the improvement in mechanical properties for the same percentage addition.

The effect of adding a third or fourth metal to a binary lead alloy can be too complex to predict. However, one indication of final results is the effect of adding the first alloying metal. In binary, ternary, and higher alloys, there are two other factors, besides a diminishing rate of mechanical improvement and increasing brittleness, which keep the optimal alloy concentrations low. These are cost and the loss of corrosion resistance. Silver, for example, improves corrosion resistance of lead in electrolytic processes, but its cost militates against broad use. Lithium, on the other hand, can make lead more vulnerable to corrosion in electrolytic processes. The amount of lithium that can be added to lead, therefore, can be limited by the unfavorable change in corrosion resistance.

The alloys of lead most frequently used in applications requiring high corrosion resistance are those which contain antimony, silver, tin, calcium, and copper. The addition of 1–13% antimony to lead creates alloys that have much greater tensile strength, resistance to fatigue, and hardness than pure lead. This is why antimonial lead is often called hard lead and the removal of antimony in lead refining is called softening. Most antimonial lead is used for storage-battery components, and contains 3–8% antimony. Alloys with antimony levels of 0.80–1.15% and 6–8% are used most frequently in nonbattery applications. The low-end alloys are used to produce cable sheathing, while the 6–8% antimony alloys are used to fabricate a wide variety of equipment, such as tank linings, pipe, and anodes for chromium plating. Lead alloys with higher percentages of antimony are used to make castings when hardness is important. However, such alloys are rarely used because they are very brittle and do not have high corrosion resistance.

The addition of small amounts of calcium and tin to lead creates alloys which have significantly increased mechanical strength and which age-harden at room temperatures. The corrosion resistance of these alloys also is higher than that of antimony alloys in many applications. Calcium alloys, with or without tin, are used as battery grids, anodes, and roofing materials.

Tin alloys of lead have much lower densities and melting points than pure lead. These properties continue to decrease as more tin is added until the eutectic composition (approximately 63% Sn, 37% Pb) is reached. The lower melting point of these lead-tin alloys and increased strength have made them especially important as solders. The alloys used primarily for corrosion resistance, however, are low in tin content, with the most important alloy, terne, usually having a composition of 12–20% Sn. Terne is used to coat steel for use in items like roofing and automotive fuel tanks. *See* CORROSION; LEAD; LEAD ALLOYS; METAL, MECHANICAL PROPERTIES OF. A. L. Ponikvar

Bibliography. *ASM Metals Handbook*, vol. 11, pp. 493–522, 9th ed., 1979; D. H. Beilstein, The Herculaneum lead smelter of St. Joe Minerals Corporation, *Proceedings of the AIME World Symposium on Mining and Metallurgy*, *Lead and Zinc*, *St. Louis*, 1970; T. R. A. Davey and W. R. Bull, Process research on lead and zinc extraction, *Proceedings of the AIME World Symposium on Mining and Metallurgy*, *Lead and Zinc*, *St. Louis*, 1970; W. Hofman, *Lead and Lead Alloys*, 1970; M. L. Jaeck (ed.), *International Symposium on Primary and Secondary Lead Processing*, vol. 15, 1989; *Lead for Corrosion Resistant Applications*, Lead Industries Association, Inc., 1974; *A Metal for the Future*, St. Joe Minerals Corp., 1975; D. O. Rausch et al. (eds.), *Lead-Zinc Update*, 1977.

Leaf

A lateral appendage which is borne on a plant stem at a node (joint) and which usually has a bud in its axil. In most plants, leaves are flattened in form, although they may be nearly cylindrical with a sheathing base as in onion. Leaves usually contain chlorophyll and are the principal organs in which the important processes of photosynthesis and transpiration occur.

Morphology

A complete dicotyledon leaf consists of three parts: the expanded portion or blade; the petiole which supports the blades; and the leaf base. Stipules are small appendages that arise as outgrowths of the leaf base and are attached at the base of the petiole (**Fig. 1**). Stipules may be green and bladelike, as in pea; coarse rigid spines, as in black locust; sheaths, as in smart-weed; or more often, temporary structures which function only to protect the leaf while it is folded in the bud. Leaves that have a blade and petiole but no stipules are said to be exstipulate. Some leaves have no apparent petiole and are described as sessile. The leaves of monocotyledons may have a petiole and a blade, or they may be linear in shape without differentiation into these parts; in either case the leaf base usually encircles the stem. The leaves of grasses consist of a linear blade attached to the stem by an encircling sheath. At the junction of the sheath and the blade is a collarlike structure called a ligule.

Arrangement (phyllotaxy). Leaves are borne on a stem in a definite fixed order, or phyllotaxy, according to species (**Fig. 2**). Phyllotaxy is usually helical (spiral), with a single leaf at each node. Leaves may be borne in a compact arrangement as in rosette plants, or if the stem between the nodes (the internode) elongates so that successive leaves are widely separated, the plant is referred to as caulescent. If two leaves are borne on each node, the arrangement is opposite; if successive pairs of opposite leaves are

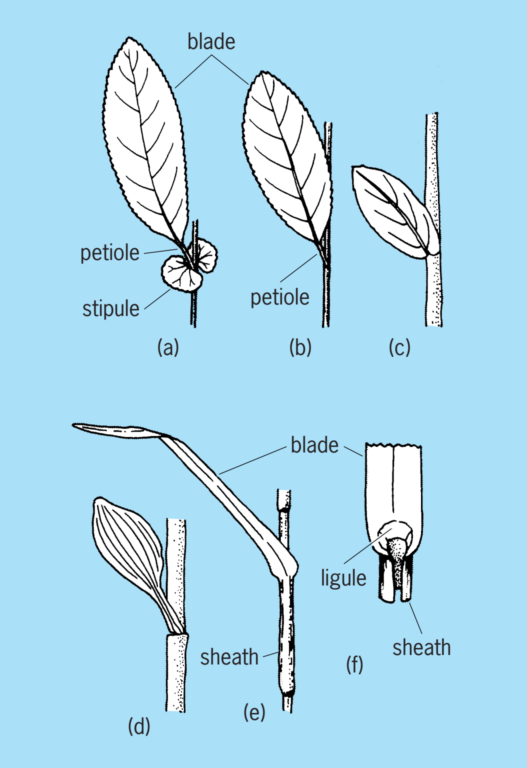


Fig. 1. Leaf parts in different structural patterns. (*a*) Complete (stipulate and petiolate). (*b*) Exstipulate. (*c*) Sessile. (*d*) Expanded monocotyledon leaf with sheathing base. (*e*) Grass leaf. (*f*) Detail of *e*.

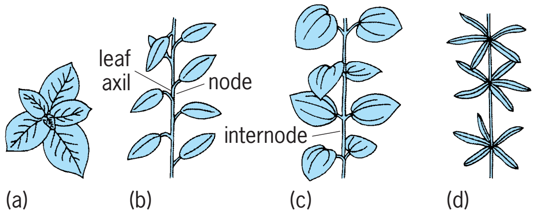


Fig. 2. Leaf arrangement. (*a*) Helical (top view). (*b*) Helical with elongated internodes (alternate). (*c*) Opposite (decussate). (*d*) Whorled (verticillate).

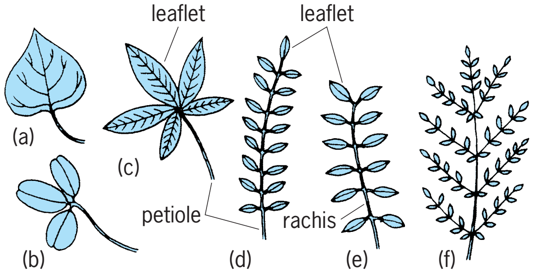


Fig. 3. Leaf types. (*a*) Simple. (*b*) Trifoliate. (*c*) Palmately compound. (*d*) Odd-pinnately compound. (*e*) Even-pinnately compound. (*f*) Decompound.

borne at right angles, it is decussate. When more than two leaves are present at a node, the arrangement is whorled (verticillate).

Types. A leaf is simple when it has only one blade. It is compound when the blade is divided into two or more separate parts called leaflets: trifoliolate if it has three leaflets (**Fig. 3**); palmately compound if leaflets originate from a common point at the end of the petiole; pinnately compound if leaflets are borne on the rachis (continuation of the petiole); odd-pinnate if such a compound leaf is terminated by a leaflet; even-pinnate if it is without a terminal leaflet; decompound (bipinnate) if it is twice compound.

Shapes. A leaf may be linear, long and narrow, with the sides parallel or nearly so (**Fig. 4**); lanceolate, narrow but tapering from base toward apex; oblanceolate, broader at the apex, tapering toward the base; spatulate, broad and obtuse at the apex, tapering to a narrow base; ovate, egg-shaped and broadest toward the base; obovate, the reverse of ovate, broadest toward the apex; elliptic, broadest at the middle and tapering slightly to a broadly rounded base and apex; oblong, somewhat rectangular, with nearly straight sides and a rounded base and apex; deltoid, triangular; reniform, kidney-shaped, broader than it is long; orbicular, circular or nearly so; peltate, shield-shaped, usually a circular leaf with the petiole attached at or near the center of the lower surface; perfoliate, having the stem apparently passing through it; connate, when the bases of two opposite leaves seem to have fused around the stem.

Margins. The margin, or edge, of a leaf may be entire (without indentation, or teeth); serrate (with sharp teeth pointing forward), or serrulate (finely serrate); dentate, with coarse teeth pointing outward, or denticulate (finely dentate); crenate or scalloped, with broad rounded teeth; undulate, with a wavy margin; incised, cut into irregular or jagged teeth or segments (if segments are narrow and pointed, laciniate; if directed backward, runcinate); pinnatifid, deeply pinnately parted (featherlike); or dissected, cut into numerous slender, irregularly branching divisions (**Fig. 5**).

When the blade is deeply cut into fairly large portions, these are called lobes. The degree of such lobing may be designated by the following terms: lobed, with sinuses usually not more than halfway from margin to midrib (midvein) or base, and with lobes and sinuses more or less rounded; cleft, when incisions extend halfway or more from margin to midrib, especially when they are sharp; parted, cut so deeply that the sinuses extend almost to the midrib or base; divided, cut entirely to the midrib, making a leaf compound.

Tips and bases. The tip of a leaf may be acuminate, gradually tapering to a sharp point; acute, tapering more abruptly to a sharp point; obtuse, with a blunt or rounded tip; truncate, seeming to be cut off square or nearly so; emarginate, decidedly notched at the

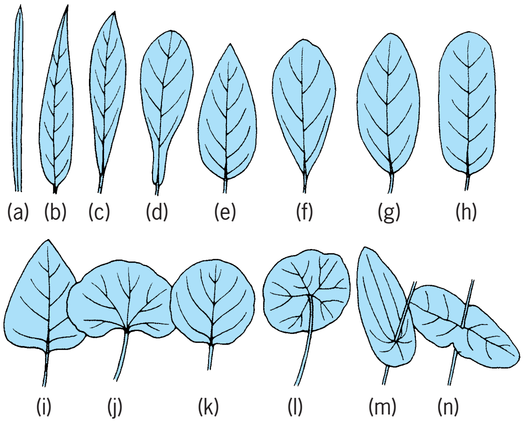


Fig. 4. Leaf shapes. (*a*) Linear. (*b*) Lanceolate. (*c*) Oblanceolate. (*d*) Spatulate. (*e*) Ovate. (*f*) Obovate. (*g*) Elliptic. (*h*) Oblong. (*i*) Deltoid. (*j*) Reniform. (*k*) Orbicular. (*l*) Peltate. (*m*) Perfoliate. (*n*) Connate.

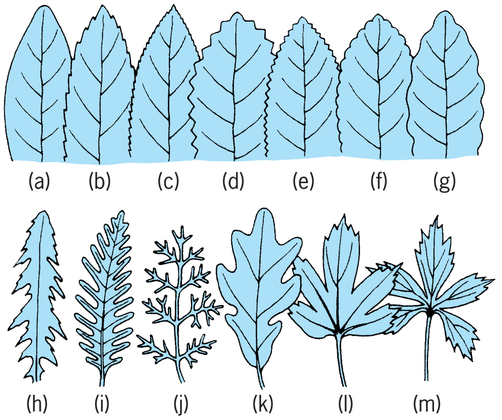


Fig. 5. Leaf margins of various types. (*a*) Entire. (*b*) Serrate. (*c*) Serrulate. (*d*) Dentate. (*e*) Denticulate. (*f*) Crenate. (*g*) Undulate. (*h*) Incised. (*i*) Pinnatifid. (*j*) Dissected. (*k*) Lobed. (*l*) Cleft. (*m*) Parted.

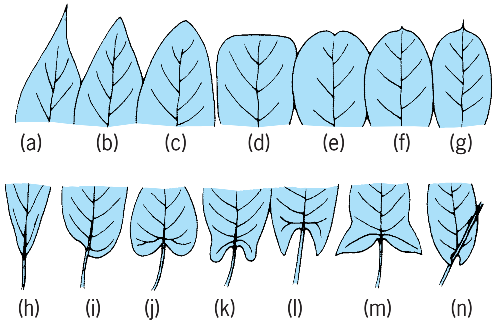


Fig. 6. Leaf tips and bases. (*a*) Acuminate. (*b*) Acute. (*c*) Obtuse. (*d*) Truncate. (*e*) Emarginate. (*f*) Mucronate. (*g*) Cuspidate. (*h*) Cuneate. (*i*) Oblique. (*j*) Cordate. (*k*) Auriculate. (*l*) Sagittate. (*m*) Hastate. (*n*) Clasping.

tip but not lobed; mucronate, abruptly tipped with a small short point; or cuspidate, ending in a sharp rigid point (**Fig. 6**).

The base of the blade may be cuneate, wedge-shaped; oblique, the two sides of the base unequal; cordate, heart-shaped, with a conspicuous sinus; auriculate, with a small earlike lobe on either side of the petiole; sagittate, arrow-shaped, with a pair of basal lobes turned outward; or clasping, sessile and partly investing the stem.

Venation. The arrangement of the veins, or vascular bundles, of a leaf is called venation (**Fig. 7**). As seen with the unaided eye, the venation appears in three basic patterns: dichotomous, reticulate (netted), and parallel. In dichotomous venation the veins are forked, with each vein dividing at intervals into smaller veins of approximately equal size; this pattern is characteristic of most ferns, gingko, and some primitive angiosperms. Reticulate venation may be described as a branching system with successively thinner veins diverging as branches from the thicker veins. The main veins may form a pinnate or palmate pattern, while the small veins form a fine network. In parallel-veined leaves, veins of relatively uniform size are oriented longitudinally, or nearly so, depending on the degree of expansion of the blade. The main longitudinal veins are usually interconnected with small veins. Reticulate venation is most common in dicotyledons, parallel venation in monocotyledons. *See* GINKGOALES; LILIOPSIDA; MAGNOLIOPSIDA; POLYPODIALES.

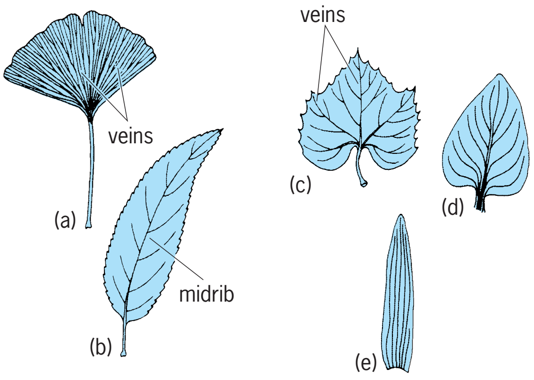


Fig. 7. Leaf venation. (*a*) Dichotomous. (*b*) Pinnate reticulate. (*c*) Palmate reticulate. (*d*) Parallel (expanded leaf). (*e*) Parallel (linear leaf).

Surface. Surfaces of leaves provide many characteristics that are used in identification. A surface is glabrous if it is smooth or free from hairs; glaucous if covered with a whitish, waxy material, or "bloom"; scabrous if rough or harsh to the touch; pubescent, a general term for surfaces that are hairy as opposed to glabrous; puberulent if covered with very fine, downlike hairs; villous if covered with long, soft, shaggy hairs; hirsute if the hairs are short, erect, and stiff; and hispid if they are dense, bristly, and harshly stiff.

Texture. The texture may be described as succulent when the leaf is fleshy and juicy; hyaline if it is thin and almost wholly transparent; chartaceous if papery and opaque but thin; scarious if thin and dry, appearing shriveled; and coriaceous if tough, thickish, and leathery.

Duration. Leaves may be fugacious, falling nearly as soon as formed; deciduous, falling at the end of the growing season; marcescent, withering at the end of the growing season but not falling until toward spring; or persistent, remaining on the stem for more than one season, the plant thus being evergreen. *See* DECIDUOUS PLANTS; EVERGREEN PLANTS.

Anatomy

The foliage leaf is the chief photosynthetic organ of most vascular plants. Although leaves vary greatly in size (from less than 0.04 in. or 1 mm in the duckweeds to over 66 ft or 20 m in some palms) and form (Figs. 1–7), they share the same basic organization of internal tissues and have similar developmental pathways. Like the stem and root, leaves consist of three basic tissue systems: the dermal tissue system, the vascular tissue system, and the ground tissue system. However, unlike stems and roots which usually have radial symmetry, the leaf blade usually shows dorsiventral symmetry, with vascular and other tissues being arranged in a flat plane.

Stems and roots have apical meristems and are thus characterized by indeterminate growth; leaves lack apical meristems, and therefore have determinate growth. Because leaves are more or less ephemeral organs and do not function in the structural support of the plant, they usually lack secondary growth and are composed largely of primary tissue only. *See* APICAL MERISTEM; ROOT (BOTANY); STEM.

The internal organization of the leaf is well adapted for its major functions of photosynthesis, gas exchange, and transpiration. The photosynthetic cells, or chlorenchyma tissue, are normally arranged in horizontal layers, which facilitates maximum interception of the Sun's radiation. The vascular tissues form an extensive network throughout the leaf so that no photosynthetic cell is far from a source of water, and carbohydrates produced by the chlorenchyma cells need travel only a short distance to reach the phloem in order to be transported out of the leaf (**Figs. 8** and **9**). The epidermal tissue forms a continuous covering over the leaf so that undue

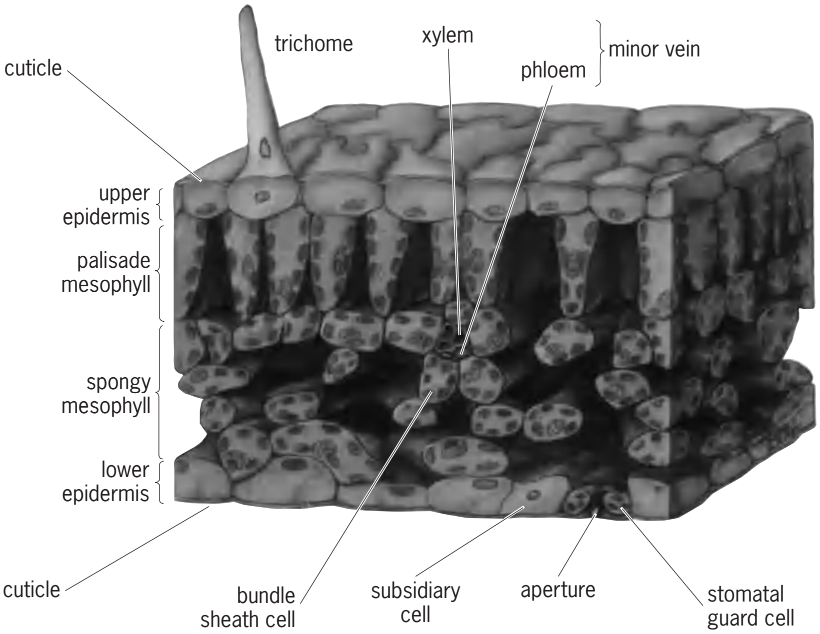


Fig. 8. Three-dimensional diagram of internal structure of a typical dicotyledon leaf.

water loss is reduced, while at the same time the exchange of carbon dioxide and oxygen is controlled.

Epidermis. The epidermis is usually made up of flat tabular cells which may be elongate, as in the linear leaves of grasses, or more or less square or lobed in surface view, as in many broad-leaved dicotyledons (Fig. 8 and **Fig. 10**). Regardless of shape, epidermal cells always fit tightly together without intercellular spaces, and they secrete a layer of hydrophobic substances, such as cutin and waxes, on the outside surface. Both of these adaptations reduce water loss. The layer of cutin is referred to as the cuticle.

The epidermal cells usually contain colorless plastids rather than chloroplasts, so that the epidermis tends to be a clear unpigmented layer of cells which allows light to penetrate to the subjacent photosynthetic tissue. The gas exchange required for photosynthesis takes place through thousands of minute stomates which usually occur at densities of about 0.15–0.75 in.^{2} (100–500/mm^2) [Fig. 10]. Each stomatal apparatus consists of a pore surrounded by two guard cells and associated subsidiary cells, which are sometimes morphologically differentiated from other epidermal cells. In the vast majority of vascular plants, the stomatal pore is closed at night so that evaporation of water vapor from internal air spaces is reduced, while in the light the guard cells are deformed in such a way that the pore is opened and carbon dioxide may enter the leaf. The ingenious mechanism of stomate opening and closure is as yet incompletely understood, but it appears to involve the pumping of potassium ions from the subsidiary cells to the guard cells, raising the osmotic concentration and increasing the hydrostatic pressure within the guard cells. The circumferential arrangement of cellulose microfibrils in the guard cell wall allows longitudinal expansion only; in addition, the ends of the guard cells are fixed, so that with increase in pressure the cells bend away from each other, opening the pore. The guard cells are unique in that they do possess chloroplasts, which provide a ready source of energy for this process. Stomates may occur on both the upper and lower leaf surfaces; or they may be restricted to the lower, shaded epidermis. In some cases such as the floating leaves of the water lily, stomates are present only in the upper epidermis.

The leaf epidermis of many vascular plants bears specialized hairlike structures called trichomes. These range from the simple prickle hairs on the leaves of many grasses, to the multicellular glandular hairs which give plants like geranium and tomato their characteristic odor, and to the complex highly branched hairs of plants like the wooly mullein. Except in obvious cases such as the modified trichomes of Venus' flytrap which secrete digestive

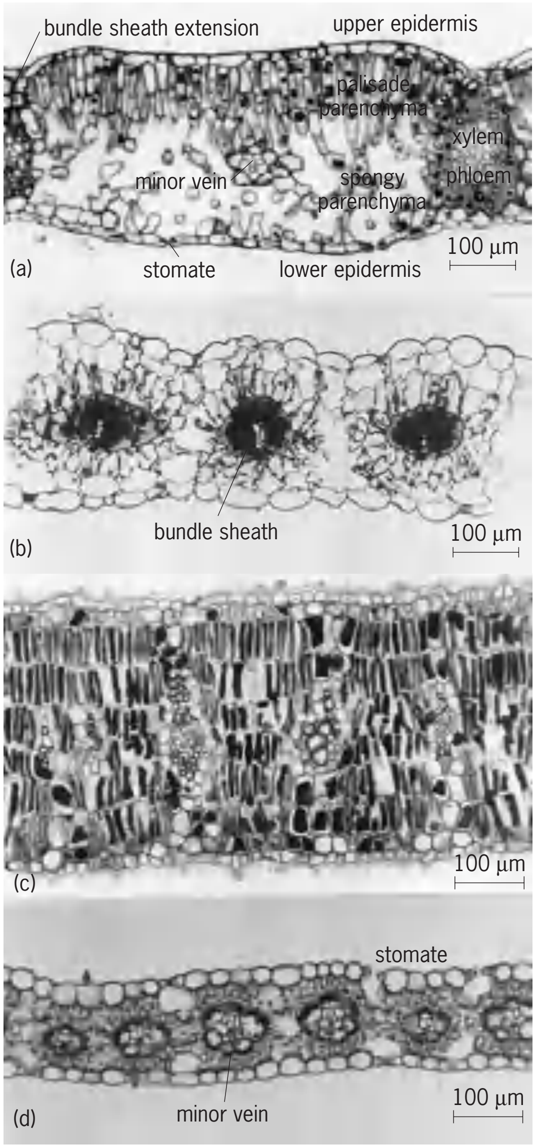


Fig. 9. Leaf cross sections. (*a*) Beech, *Fagus grandifolia*. (*b*) *Atriflex rosea*. (*c*) *Eucalyptus* sp. (*d*) Corn, *Zea mays*.

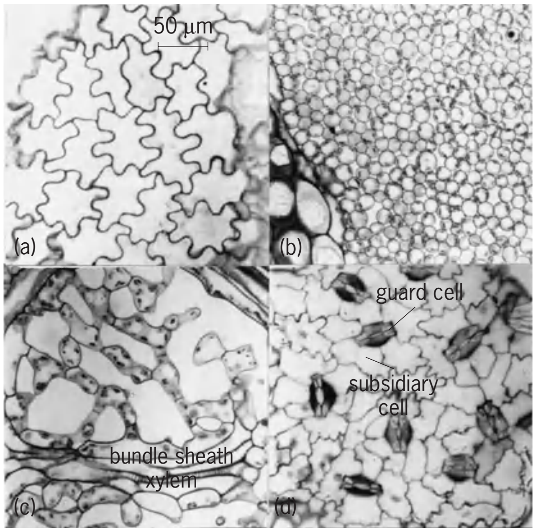


Fig. 10. Sections of beech (*Fagus grandifolia*) cut parallel to the surface (paradermal sections). (*a*) Upper epidermis. (*b*) Palisade parenchyma. (*c*) Spongy parenchyma. (*d*) Lower epidermis. (*From N. G. Dengler, L. B. MacKay, and L. M. Gregory, Cell enlargement and tissue differentiation during leaf expansion in beech, Fagus grandifolia, Can. J. Bot., 53:2846–2865, 1975*)

enzymes, the actual function of these widespread trichomes has not been demonstrated experimentally. Trichomes are believed to reduce water loss through the cuticle and stomates by creating a boundary layer on the leaf surface. A dense layer of trichomes may also reflect solar radiation from the leaf, protecting internal enzymes from damaging high temperature. In some cases trichomes may prevent insects, snails, and slugs from eating the leaves. *See* EPIDERMIS (PLANT).

Mesophyll. The ground tissue of leaves is referred to as mesophyll and consists of loosely packed parenchyma cells separated by intercellular spaces (Figs. 8 and 9). Within these cells are numerous green chlorophyll-containing chloroplasts, which are the site of photosynthesis. The mesophyll is frequently differentiated into palisade parenchyma and spongy parenchyma. The palisade parenchyma consists of one or more layers of columnar cells elongated at right angles to the epidermal layer, while the spongy parenchyma is made up of highly lobed cells with an extensive network of intercellular air spaces. Often, the largest intercellular spaces in the leaf occur in the mesophyll adjacent to a stomate, and are called substomatal chambers. In leaves with a horizontal orientation, the palisade cells usually lie next to the upper epidermis, while in leaves with a vertical orientation, layers of palisade may be adjacent to both epidermal layers, or as in the leaves of many grasses, there may be no differentiation into palisade and spongy parenchyma (Fig. 9). *See* PARENCHYMA.

Vascular tissue. The major veins of the leaf provide the connection between the vascular bundles of the stem and petiole and the xylem and phloem of the minor veins embedded in the mesophyll of the blade. Each major vein is associated with a rib of parenchymatous tissue, which usually projects from the lower surface of the leaf and can be seen with the naked eye. Strengthening tissues such as collenchyma and sclerenchyma are often associated with the major vein and provide the chief means of support for the leaf blade. Regardless of the pattern of the major veins (Fig. 6), the distribution of minor veins is such that every photosynthetic mesophyll cell is usually less than 100 micrometers away from the conducting tissues of the minor veins. Xylem occurs on the adaxial side of the minor veins and consists of both parenchyma and tracheary elements, which often have helical secondary walls (Fig. 10). Phloem occurs on the abaxial side of the vein and consists of sieve tube elements and phloem parenchyma. The conducting tissues at the minor veins are enclosed in a continuous sheath of ground tissue, the bundle sheath. The bundle sheath may be sclerenchymatous, but it more often consists of parenchyma cells with poorly developed chloroplasts. In many leaves the minor veins are connected with the epidermis by panels of bundle sheath cells called bundle sheath extensions. In leaves with freely ending veinlets, these extensions pass over the end of the veinlets so that the vascular tissue is not exposed directly to the intercellular spaces. In some plants, guttation (secretion of water) takes place from specialized vein endings and associated tissues called hydathodes. *See* PHLOEM; XYLEM.

Internal structure in relation to transport. Transpiration water leaving the tracheary elements must cross the bundle sheath parenchyma to reach the photosynthetic mesophyll. The bulk of the water is believed to travel either through the apoplast, that is, outside the living protoplast, or through the walls of the intervening cells. The evaporation of water from the cell walls of the mesophyll into the intercellular spaces provides the driving force for the bulk movement of the continuous column of water that extends from the soil through the xylem tracheary elements of root, stem, and leaf to the mesophyll cell walls. It is the cohesive force of this column which holds all of these together. In addition to moving directly from the xylem to the mesophyll, water moves from the veins through the bundle sheath extensions and then laterally along the epidermis. This pathway is particularly important for the palisade parenchyma cells which have limited lateral contact with one another. The bulk of the water lost from the leaf by transpiration evaporates from the surface of the internal tissues of the leaf, often 10 to 30 times the area of external leaf surface. Water vapor escapes from the leaves through stomatal pores and, to a much lesser extent, through the cuticle. *See* PLANT-WATER RELATIONS.

The photosynthetic mesophyll cells act as the source of carbohydrates for the rest of the plant. Sugars are believed to move in a symplastic pathway, a pathway in which materials move from cell to cell through cytoplasmic connections called plasmodesmata. Materials following this pathway do not leave the living protoplast as they move from the mesophyll, across the bundle sheath, and into the sieve tube elements of the phloem. The process of

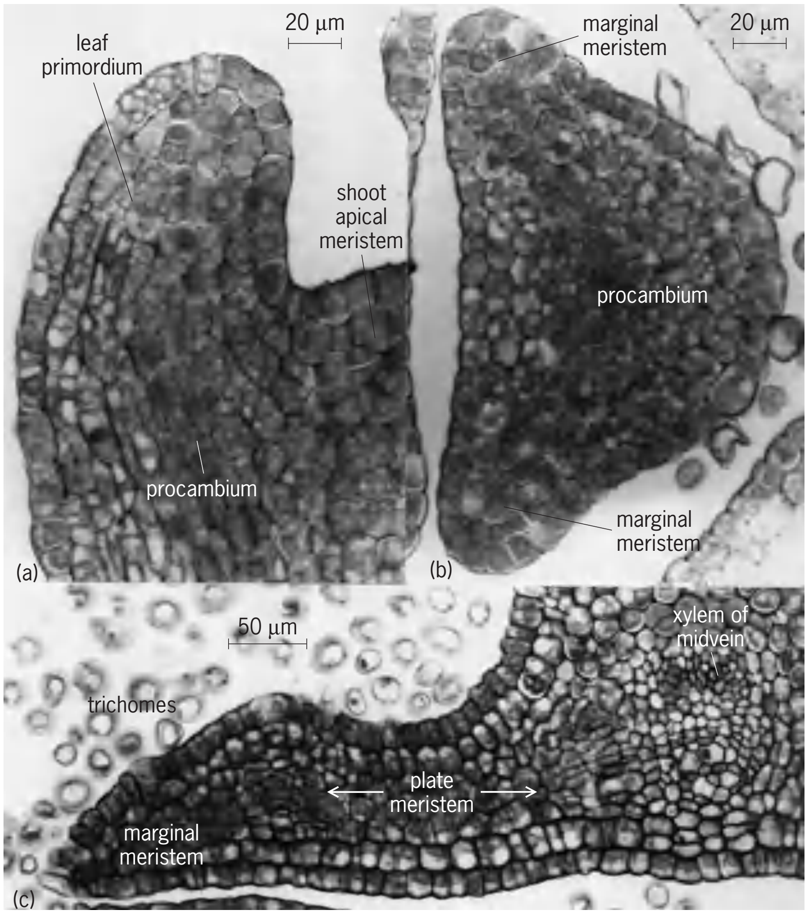


Fig. 11. Sections of developing leaves of *Calycanthus occidentalis*. (*a*) Longitudinal section of leaf primordium. (*b*) Cross section of leaf primordium. (c) Cross section of young leaf blade. (*From N. G. Dengler, Ontogeny of the vegetative and floral apex of Calycanthus occidentalis, Can. J. Bot., 50:1349–1356, 1972*)

loading carbohydrates against a concentration gradient into the sieve tube elements requires metabolic energy. The parenchyma cells surrounding the conducting cells often display features of transfer cells—cells with extensive wall invaginations which vastly add to the surface area of the cell membrane. *See* PLANT TRANSPORT OF SOLUTES.

C4 plants. In most monocotyledonous and dicotyledonous species, a three-carbon compound, 3-phosphoglyceric acid, is the first product of photosynthesis. These are called C3 plants. However, since the discovery of a new photosynthetic pathway in the 1960s, a growing number of plants have been shown to initially produce four-carbon organic acids as the first products of photosynthesis. These C4 plants are normally characterized by a distinctive type of leaf anatomy in which the bundle sheath parenchyma is made up of conspicuous dark green cells containing a number of large chloroplasts localized at one end of the cell (Fig. 9*b* and *d*). In addition, the photosynthetic mesophyll cells are often arranged concentrically around the bundle sheath in C4 plants, giving rise to the designation Kranz (wreath) anatomy for these species. Plants with the C4 pathway are more efficient in their incorporation of carbon dioxide than C3 plants, particularly at high light intensities and high temperatures. *See* PHOTOSYNTHESIS.

Development

The foliage leaf, regardless of size and shape at maturity, follows the same basic sequence of developmental events in its formation.

Initiation. The leaf arises as an emergence on the flanks of the apical meristem in the shoot tip. Localized cell divisions and accompanying cell enlargement first give rise to a small mound of cells and then to a fingerlike projection called a leaf primordium (**Fig. 11*a***). The new leaf primordium acts as a growth center, drawing nutrients from surrounding tissues and producing growth regulators such as auxin. Microsurgical experiments have shown that the growth center induces the differentiation of a strand of future vascular tissue called procambium which is in continuity with the vascular system of the stem. The procambium continues to grow toward the tip of the primordium as the primordium itself grows in length. The leaf primordium axis later forms the petiole and midrib of the mature foliage leaf, and the first procambial strand becomes the midvein of the leaf. Initially the leaf primordium elongates through concentrated meristematic activity near the apex; eventually, cells at the primordium apex begin to mature, and further growth in length is the result of cell divisions scattered throughout the young leaf. *See* BUD.

Blade formation. The leaf blade is usually initiated when the leaf primordium is less than a half millimeter in length. The blade arises as a ridge of tissue produced by a zone of cell division and cell enlargement on either side of the primordial axis called marginal meristems (Fig. 11*b*). Analysis of the planes of divisions in the marginal meristems has shown that cell divisions in the surface layer are largely perpendicular to the leaf surface, giving rise to a continuous sheet of cells which will later mature as epidermis. Cell divisions in the internal tissues occur in a variety of planes, adding to the volume of the developing leaf blade, and determining the number of cell layers in the mature leaf. Cell divisions in both the surface and internal layers are most frequent in a region several cells removed from the actual leaf margin. In many plant species, the marginal meristems are active for only a limited period of time, and expansion of the leaf blade is the result of cell division and enlargement throughout the blade between the margin and midrib called plate meristem activity. The plane of cell divisions in the plate meristem is predominantly perpendicular to the leaf surface. This means that the number of cell layers established by the marginal meristem is perpetuated, and that with cell enlargement, the leaf expands in size manyfold in a horizontal plane.

At the time that the leaf expands from the bud, between 50 and 90% of its cells have already been formed. Therefore, the final stages of leaf expansion are mostly the result of enlargement of cells that have been formed by divisions in the plate meristem. Characteristic differences in the rate, duration,

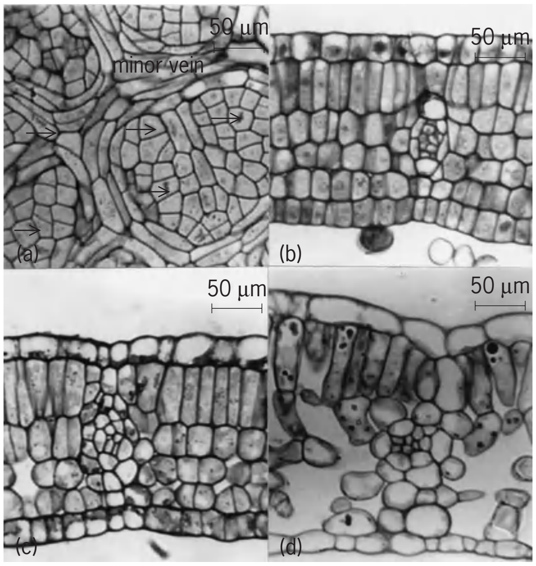


Fig. 12. Sections of beech (*Fagus grandifolia*) leaves during expansion from the winter bud. (*a*) Paradermal section of spongy parenchyma at time of bud swelling. Arrows indicate cell divisions in plate meristem. (*b*) Cross section of leaf at same stage. (*c*) Cross section of leaf 1 week later. (*d*) Cross section of leaf 2 weeks later. (*From N. G. Dengler, L. B. MacKay, and L. M. Gregory, Cell enlargement and tissue differentiation during leaf expansion in beech, Fagus grandifolia, Can. J. Bot., 53:2846–2865, 1975*)

and direction of cell enlargement occur between the various tissues of the leaf blade. For instance, cells that will become the palisade parenchyma elongate greatly in a vertical direction, and very little in a horizontal direction (**Fig. 12**); in addition, cell divisions occur for a longer period in the palisade layers. In contrast, cells of the upper epidermis undergo a great deal of horizontal expansion and very little vertical expansion. The intercellular spaces of the palisade parenchyma layers are formed when the enlargement of the upper epidermal cells draws the attached palisade cells apart. Cell expansion in the spongy parenchyma layers does not occur uniformly over the cell surface, resulting in a lobed cell shape and an extensive network of intercellular spaces. The procambial strands that will differentiate as minor veins are formed from localized cell divisions in the plate meristem (Fig. 12). The minor vein pattern is blocked out early during plate meristem activity, and the areoles, regions of mesophyll surrounded by vascular tissue, are well defined throughout the period of leaf expansion. Maturation of xylem and phloem occurs first in the major veins, starting at the base of the leaf and proceeding toward the apex. This is followed by maturation of the xylem and phloem of the minor veins, which proceeds from the apical portion of the leaf blade toward the base. *See* PLANT GROWTH.

Variations in developmental pattern. The process of leaf development in plants with lobed or compound leaves does not differ qualitatively from the mechanisms described above. Rather, differences in leaf shape arise through differences in the location and direction of marginal and plate meristem activity. The long linear leaves of grasses and some other monocotyledons, however, do possess some unique developmental features. At the earliest stages of development, the leaf primordium is crescent-shaped and encircles the shoot apical meristem. After a short period of apical growth along the rim of the crescent, most of the meristematic activity is localized at the base of the leaf in the intercalary meristem. A majority of cell divisions are in a plane which is perpendicular to the future long axis of the leaf. This means that cells of all tissue layers are formed in long files, with those cells near the apex of the leaf reaching maturity while new cells are being added to the base of each file. Intercalary growth may be prolonged, and may occur sporadically throughout the life of an individual leaf. This unique pattern of growth is often regarded as an adaptation which is associated with the evolution of grazing mammals.

Leaf development in rapidly growing herbaceous plants is a continuous process in which leaves produced on the shoot apical meristem expand, carry on photosynthesis, and eventually senesce when the reproductive stage of the life cycle nears an end. In most perennial plants such as the beech illustrated in Figs. 9, 10, and 12, new leaf primordia are formed late in the growing season and overwinter in the apical bud. In the spring the young leaves can quickly complete plate meristem activity and leaf expansion so as to maximize the period of photosynthesis. Leaf abscission occurs at the end of the growing season in deciduous woody plants. The abscission zone may be structurally weak, but the actual separation of leaf from stem is the result of enzymatic degradation of cell walls. A protective layer is produced by the deposition of protective substances such as suberin and wound gum, and the wound is often further protected by the development of a layer of cork. The resulting leaf scar and the bundle scars which mark the position of the abscission of the leaf midvein are often distinguishing features in the identification of species. *See* ABSCISSION; PERIDERM; PLANT ANATOMY.

Nancy G. Dengler

Lean manufacturing

A unique linked-cell manufacturing system. Initiated in the 1960s by the Toyota Motor Company, it is also known as the Toyota Production System (TPS), the Just-in-Time/Total Quality Control (JIT/TQC) system, or World Class Manufacturing (WCM) system. In 1990, it was given a name that would become universal, "lean production." This term was coined by John Krafcik, an engineer working in the International Motor Vehicle program at MIT with J. P. Womack, D. Roos, and D. T. Jones, who observed that this new system used less of the key resources needed to make goods. What is different about the system is its use of manufacturing cells linked together with a functionally integrated system for inventory and production control. The result is low cost (high efficiency), superior quality, and on-time delivery of unique products from a flexible system. *See* FLEXIBLE MANUFACTURING

SYSTEM; INVENTORY CONTROL; MANUFACTURING ENGINEERING.

Mass versus lean production. The key proprietary aspects of lean production are U-shaped manufacturing and assembly cells, using walking workers and are designed with system requirements in mind. Design decoupling allows the separation (decoupling) of processing times for individual machines from the cycle time for the cell as a whole, permitting the lead time to make a batch of parts to be independent of the processing times for individual machines. This takes all the variation out of the supply chain lead times, so scheduling of the supply system can be greatly simplified. The supply chain is controlled by a pull system of production control called Kanban. Using Kanban, the inventory levels can be dropped, which decreases the throughput time for the manufacturing system. *See* SUPPLY CHAIN MANAGEMENT.

Figure 1 shows an entire mass production factory reconfigured into a lean manufacturing factory. The final assembly lines are converted to mixed model final assembly; this levels the demand for subassemblies and other components. The rate of production is determined by recalculating the monthly demand into a daily demand and trying to make the same product mix every day.

The subassembly lines are reconfigured into U-shaped cells. The daily output from these cells is balanced to match the demand from final assembly.

The traditional job shop generates a variety of unique products in low numbers with its functional design. It is redesigned into U-shaped manufacturing cells that produce families of component parts (**Fig. 2**).

Lean manufacturers focus on sole-sourcing each component or subassembly (that is, they do not have multiple vendors supplying the same components), sharing their knowledge and experience in linked-cell manufacturing with their vendors on a one-to-one basis. For lean automobile manufacturers, the final assembly plant may have only 100 to 400 suppliers, with each supplier becoming a lean or JIT vendor to the company. In the future, the number of vendors supplying a lean manufacturer will decrease even more. The Mercedes-Benz plant in Vance, Alabama, has around 80 suppliers. The subassemblies contain more components as the vendors take on more responsibility for the on-time delivery of a larger portion of the auto.

Lean cell design. In a true lean manufacturing system, manufacturing processes and equipment are designed, built, tested, and implemented into the manufacturing cells. The machine tools and processes, the tooling (workholders, cutting tools), and the material-handling devices (decouplers) are designed specifically for cellular manufacturing. Simple, reliable equipment that can be easily maintained should be specified. In general, flexible, dedicated equipment can be built in-house better than it can be purchased and modified for the needs of the cell.

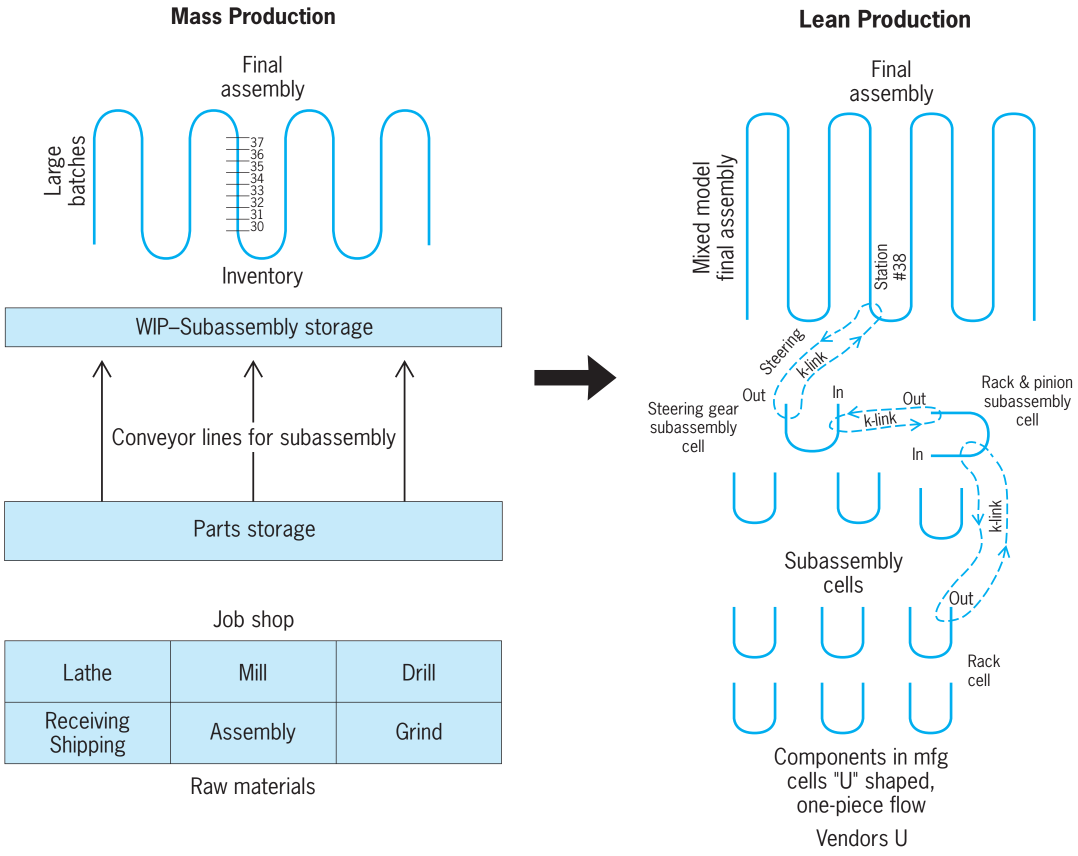


Fig. 1. Restructuring of a job shop/flow shop final assembly design into a linked-cell manufacturing system.

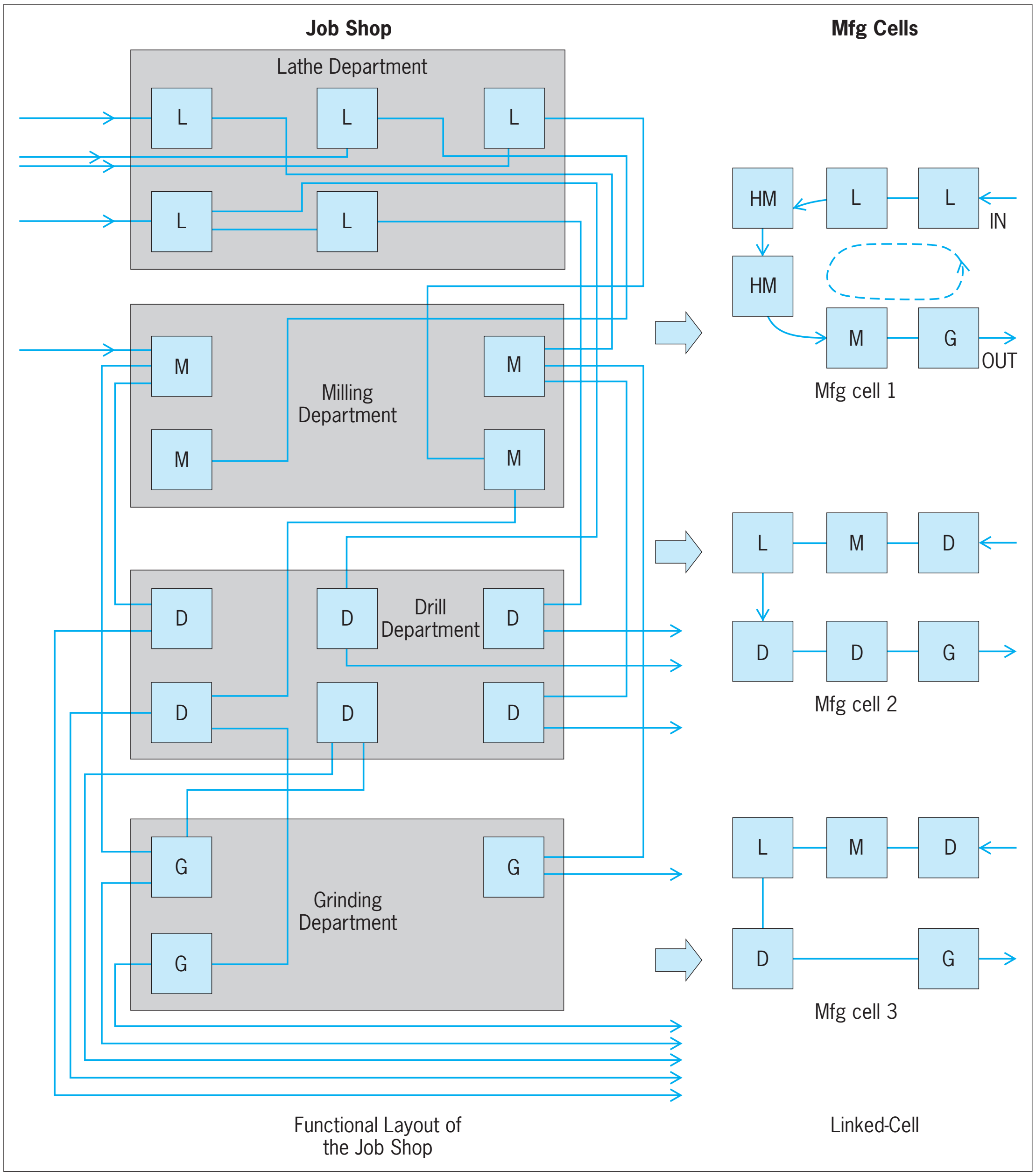


Fig. 2. Job shop portions of the factory require a system level conversion to redesign them into manufacturing cells.

However, many plants that lack the expertise to build machines from scratch do have the expertise to modify equipment to give it unique capabilities, and this interim cell approach can lead a company into true lean manufacturing cells. *See* MATERIALS-HANDLING EQUIPMENT.

Many companies understand that it is not good strategy simply to imitate manufacturing process technology from another company and then expect to make an exceptional product. When process technology is purchased from outside vendors, any unique aspects will be quickly lost. Companies must carry out research and development on manufacturing technologies and systems in order to produce effective and cost-efficient products, but the result makes such investment pay off. There are many advantages to this home-built equipment strategy.

Flexibility (process and tooling adaptable to many types of products). Flexibility requires rapid changeover of jigs, fixtures, and tooling for existing products and rapid modification for new designs. The processes have excess capacity; they can run faster if necessary, but they are designed for less than full-capacity operation.

Building to need. There are three aspects to this. First, there is no unused capability or options. Second, the machine can have unique capabilities that competitors do not have and cannot get access to through equipment vendors. For example, in the lean cell shown in **Fig. 3** is a broaching (cutting) machine for producing the gear teeth on a rack bar. The angle that these teeth make with the bar varies for different types of racks. Job shop broaching machines are not acceptable for this cell because of their

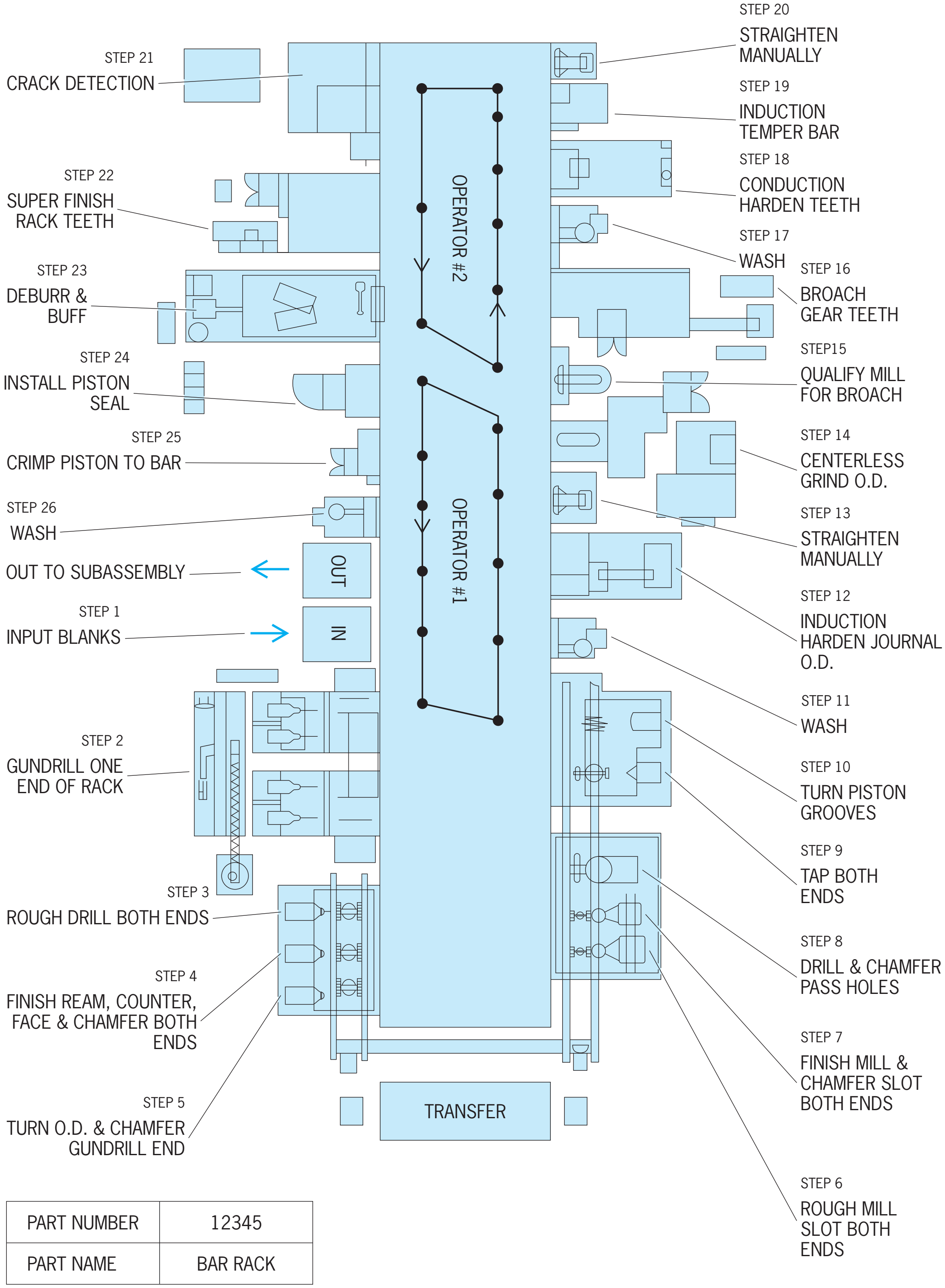


Fig. 3. Lean manufacturing cell for producing rack bars for a rack-and-pinion steering gear.

long changeover times, so a unique machine tool for broaching was designed and built using a proprietary process. Third, the equipment should allow the operator to stand and walk. Equipment should be the appropriate height to allow the operators to easily perform tasks standing up and then move to the next machine in a step or two (that is, the design should have a narrow footprint).

Built-in maintainability/reliability/durability. Equipment should be easy to maintain (to oil, clean, changeover, or replace worn parts, and to use standardized screws). Many of the cells at lean vendors are clones of each other. The vendor company, being the sole source, has the volume and the expertise to get business from many companies, making essentially the same components or subassemblies for many original equipment manufacturers (OEMs). The equipment can be interchanged from one cell to another in emergencies.

Machines, handling equipment, and tooling built for the needs of the cell and the system. Machines are typically single-cycle automatics but may have capacity for process delay. An example of process delay is a heat treatment process that takes 4 min in a cell with a 1 min cycle/time. If the heat treatment machine can hold four units, each getting 4 min of treatment, one unit is still produced per minute.

Safety. Equipment is designed to prevent accidents (fail-safe).

Equipment designed to be easy to operate, load, and unload. Research has shown that manufacturing cells have ergonomic benefits over the job shop. Toyota ergonomists recommend unloading with the left hand, loading with the right hand, while walking right to left. The wide variety of tasks over the cell cycle time keep the operators from developing cumulative trauma injuries.

Equipment designed to process single units, not batches. Small-area, low-cost equipment is the best. Machining or processing time (MT) should be modified so that it is less than the cycle time (CT), the time in which one unit must be produced. Equipment processing speed should be set in view of the cycle time, such that MT < CT. The MT is related to the machine parameters selected. For example, in a cutting operation this approach often permits the reduction of the cutting speed, thereby increasing the tool life and reducing downtime for tool changes. This approach also reduces equipment stoppages, lengthens the life of the equipment, and may improve quality.

Equipment that contains self-inspection devices (such as sensors and counters) to promote autonomation. Autonomation is the autonomous control of quantity and quality. Often the machine is equipped to count the number of items produced and the number of defects. *See* QUALITY CONTROL.

Equipment that is movable. Machines are equipped with casters or wheels, flexible pipes, and flexible wiring. There are no fixed conveyor lines.

Equipment that is self-cleaning. Equipment disposes of its own chips and trash.

Equipment that is profitable at any production volume. Equipment that needs millions of units to be profitable (R. J. Schonberger calls them supermachines) should be avoided, because if the production volume ever exceeds the maximum capacity that the first supermachine can build, it will be necessary to purchase another supermachine and the new one will not be profitable until it approaches full use. Schedulers of the equipment will typically divide the volume into two machines, so neither will be profitable.

Lean cell operation. The manufacturing cell shown in Fig. 3 typically uses two operators. These standing, walking workers move from machine to machine in counterclockwise (CCW) loops, each completing the loop in about 1 min. Operator 1 typically addresses 10 stations, and operator 2 addresses 11 stations. Most of the steps involve unloading a machine, loading another part into the machine, checking the part unloaded, and dropping the part into the decoupler (handling) elements between the machines. The stock-on-hand (SOH) in the decouplers and the machines helps to maintain the smooth flow of the parts through the machines. The decouplers also can be designed to perform inspections for part quality or necessary process delays while the parts heat up, cool down, cure, and so on. The stock-on-hand is kept as low as possible.

Sometimes the decoupler elements perform the inspection of the part, but mostly they serve to transport parts from one process to the next. The decoupler may also perform a secondary operation such as deburring or degaussing (removal of residual magnetic fields).

By design, one operator controls both the input and output of the cell. One operator always controls the volume of material going through the cell. This keeps the stock-on-hand quantity constant and keeps the cell working in balance with the final or subassembly lines it is feeding.

At the interface between the two operators, either one can perform the necessary operations, depending on when they arrive and when the processes in the machines are finished. That is, the region where the two operators meet is not fixed, but changes or shifts depending upon the way parts are moving about the cell. This is called the relay zone. This added flexibility requires that the workers are cross-trained on all the processes in the cell.

Ergonomics of lean cells. Ergonomics deals with the mental, physical, and social requirements of the job, and how the job is designed (or modified) to accommodate human limitations. For example, the machines in the cell are designed to a common height to minimize lifting of parts, transfer devices are designed for slide on/slide off, and automatic steps equipped with interrupt signaling help workers monitor the process. When the job is primarily loading/unloading, ergonomic concerns regarding lifting and placing parts in machines and operating workholding devices must be addressed. Lean manufacturing cells are relatively free of cumulative trauma syndrome (CTS) problems because the operators' tasks and movements are so varied from machine to machine. Cell designers should try to incorporate ergonomic issues initially rather than trying to implement fixes later.

Human performance in detecting and correcting cell malfunctions will establish utilization and production efficiency. The design of machines for maintainability and diagnostics is critical. In manufacturing and assembly cells where workers operate machines, it is important that all the machines are ergonomically identical. Sewing machines in a cell are a good example: To the operator, all the machines should feel the same in terms of control.

Manufacturing system integration. Many believe that the only way in which manufacturing companies can compete is to automate. This approach, known as computer integrated manufacturing (CIM), was recently renamed agile manufacturing. The concept is to achieve integration through computerization and

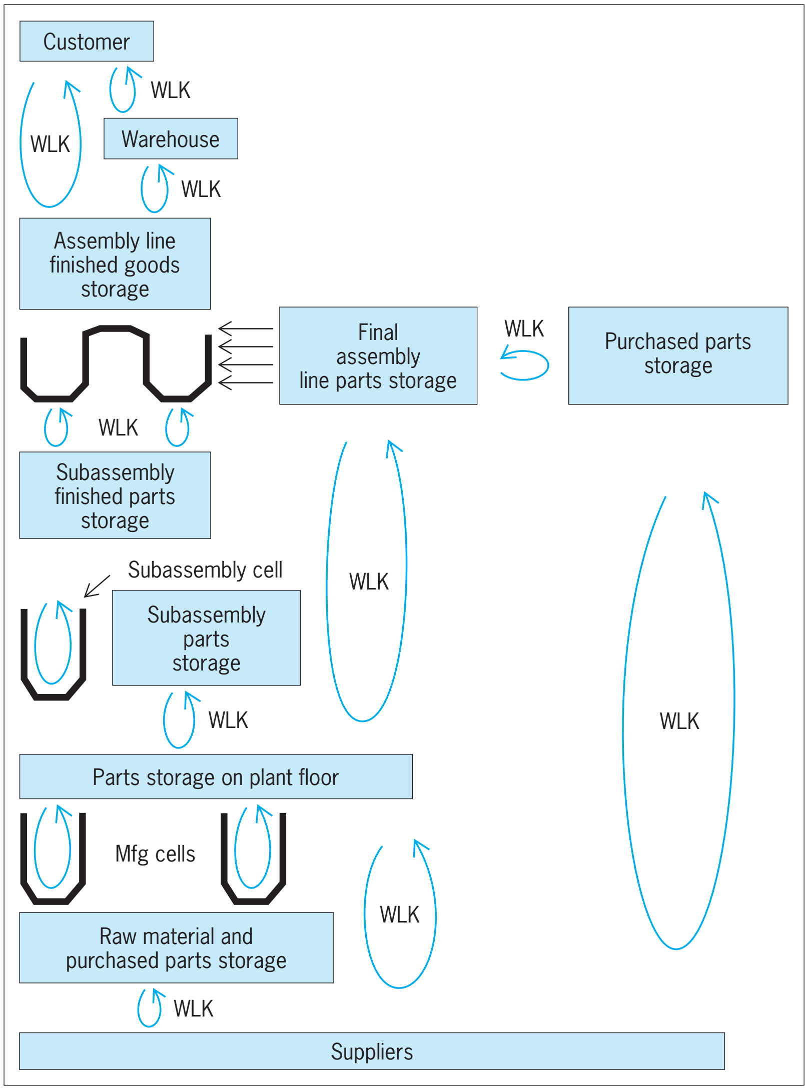


Fig. 4. Linked-cell design. In the lean manufacturing system, the manufacturing and assembly cells are linked to the final assembly area by Kanban (card) inventory links or loops. WLK indicates a withdrawal card.

automation. This often results in trying to computerize, robotize, or automate very complex manufacturing and assembly processes, and it works only when there is little or no variety in the products. Lean manufacturing systems take a different approach—integrate the manufacturing system, then computerize and automate (IM, C). Experts on CIM now agree that lean manufacturing, especially development of manufacturing and assembly cells, must come before efforts to computerize the system. While costs of these systems are difficult to obtain, the early evidence suggests that the lean cell approach is significantly cheaper than the CIM approach. *See* AUTOMATION; COMPUTER-INTEGRATED MANUFACTURING.

Outlook. The lean factory is based on a different design for the manufacturing system in which the sources of variation in time are minimized and delays in the system are systematically removed. In the linked-cell manufacturing system, in which manufacturing and assembly cells are linked together with a pull system for material and information control (**Fig. 4**), downstream processes dictate upstream production rates. The linked-cell manufacturing system strategy simplifies the manufacturing system, integrates the critical control functions before applying technology (automation, robotization, and computerization), avoids risks, and makes automation easier to implement. This is the strategy that will predominate in the next generation of factories worldwide. *See* INDUSTRIAL ENGINEERING; PRODUCTION ENGINEERING; PRODUCTION METHODS.

J T. Black

Bibliography. R. U. Ayres and D. C. Butcher, The flexible factory revisited, *Amer. Scientist*, 81:448–459, 1993; J T. Black, Cell design for lean manufacturing, *Trans. NAMRI/SME*, 23:353–358, 2000; J T. Black and B. J. Schroer, Simulation of an apparel assembly cell with walking workers and decouplers, *J. Manuf. Sys.*, 12(2):170–180, 1993; J T. Black, C. C. Jiang, and G. J. Wiens, Design, analysis and control of manufacturing cells, *PED*, vol. 53, ASME, 1991; Y. Monden, *Toyota Production System*, Industrial Engineering and Management Press, IIE, 1983; S. Nakajima, *TPM, Introduction to TPM: Total Productive Maintenance*, Productivity Press, 1988; R. J. Schonberger, *Japanese Manufacturing Techniques: Nine Hidden Lessons in Simplicity*, Free Press, 1982; K. Sekine, *One-Piece Flow: Cell Design for Transforming the Production Process*, Productivity Press, 1990; S. Shingo, *A Study of The Toyota Production System*, Productivity Press, 1989; N. P. Suh, Design axioms and quality control, *Robotics and CIM*, 9(4/A):367, August–October 1992; J. P. Womack, D. T. Jones, and D. Roos, *The Machine That Changed The World*, Harper Perennial, 1991.

Learning mechanisms

Those processes, activities, structural components, and alterations of the nervous system that produce the behavioral manifestations and changes due to experience and event associations (rather than those due to maturation, fatigue, and other developmental or temporary states).

Central issues. The search for a neural basis of learning has long been guided by several central issues, as described below.

Definition of learning. At the root of any search for the neural mechanism of learning is the problem of defining learning. Most simply, learning may be defined as a change in behavior. However, this definition is certainly too broad since any change in behavior could then be called learning whether it resulted from drug intoxication, sleep, or even death. It is now common to use the term behavioral plasticity for behavior changes that seem to involve at least some aspects of learning and to subdivide the definition into specific types of learning, such as motor learning, sensory learning, and associative learning. In this way, the definition of learning has been narrowed to indicate a change in behavior that is relatively permanent, that is not caused by development, maturation, fatigue, or other short-term processes, and that is due to experience or reinforcement. Hence, what is known as sensitization, a process in which a given behavior increases in intensity simply with repeated occurrences, is an example of behavioral plasticity, whereas learning to play basketball involves the more specific categories of both motor and sensory learning.

Several types of learning have been recognized, perhaps the simplest of which is associative learning, that is, learning by association of one stimulus with another through the repeated simultaneous occurrence of the two stimuli. This type of learning is typified by the dog that learns to associate the sound made when a package of dog meal is opened with the food being placed in its dish. This is an example of Pavlovian conditioning. Motor learning, on the other hand, occurs when an animal or human learns to perform some motor task in response to a given event or stimulus. Sensory learning can be thought of as that learning in which an organism is trained to respond to changes in some stimulus or to differences between stimuli. Included here are such familiar things as learning faces or tunes, learning to distinguish two shades of color, and connecting voices with faces. The last is an example of cross-modal generalization learning in which a stimulus in one sensory modality (the auditory) recalls the image of the face in another (the visual). In the human, there are other categories of learning, such as cognitive learning in which the human learns without any obvious external stimuli, but simply through the internal activity of the brain, commonly referred to as consciousness. This is the most difficult type of learning to investigate, since there is little control over the process. *See* COGNITION; CONDITIONED REFLEX.

It is important to recognize that in the search for neural substrates of learning, the type of learning must be specified, since it is evident that the term learning can apply to many different types of events in the brain. Even though all of these events may have at least some common neural substrates, they involve different areas of the brain—and probably different neural mechanisms as well. The problems encountered when defining learning have become stumbling blocks in determining which neural processes are to be studied, because there is little consensus about what is to be measured.

Relationship between learning and memory. The question of how learning and memory differ is important in designing studies of the underlying processes of learning. Learning is really a hypothetical construct that is presumed to take place when certain conditions are met, and it is the basis for subsequent changes in the behavior of an organism in response to its environment. Memory, on the other hand, is the storage of those alterations for later use. Without memory, there would be no way to determine if learning had occurred since whatever had been learned would be immediately forgotten. Memory, then, is always needed to assess learning and cannot be completely separated from it. In the same way, any assessment of processes underlying learning will also assess memory, although the two can be analyzed somewhat independently by appropriate experimental designs.

The term engram has long been applied to the memory trace of learning. This term denotes that special alteration which occurs as a consequence of learning and which is then translated into altered behavioral output at a later time. Thus, the engram is a stored representation of the learning experience and is presumably open to investigation. *See* MEMORY.

Physiological process. Since the behavior of all animals above the level of the simplest multicellular organisms is the product of a nervous system, the processes underlying those behavioral changes referred to as learning must be the result of some activity or change in the nervous system. The complex human nervous system includes both the central nervous system (the brain and spinal cord) and the peripheral nervous system (the nerves of the body and the autonomic ganglia). It is generally assumed that the changes important in learning occur in the central nervous system, with the peripheral nerves serving a transport function. Within the central nervous system, changes related to learning (the engram) could be limited to one specific region or be dispersed throughout the entire brain and spinal cord. In addition, the learning substrate could be at a subcellular level, within or between individual cells, or among groups of cells at a regional level. *See* NERVOUS SYSTEM (VERTEBRATE).

Only those processes that are known to occur in the brain can be considered candidates for the substrate of learning. For instance, much is known about the mechanics of blood flow and its effects on brain function, but they are unlikely to serve as the basis of learning. However, electrical events, alterations of neural structure, biochemical changes in or around neurons, changes in synapses, and alternations of neurons over large brain regions all seem to be candidates.

Generalizability. Initial observations of the brain's control of learning occurred mainly in humans with various types of brain damage. However, with the advent of new techniques for the study of brain processes, the study of brain function in infrahuman species has become necessary, since many of the techniques are invasive and destructive. Indeed, most current knowledge of brain processes and neural function has come from animal studies, and may or may not apply to humans. When the study subjects are mammals, any processes found there probably will apply to human brain function. However, in studies on animals such as *Aplysia*, an invertebrate sea snail, the parallelism may be more difficult to establish. Such arguments make the assumption that the only goal is to investigate how humans function, and while this is certainly the ultimate goal, it is not the only one. Identifying such processes in these organisms allows for better care of these animals both in captivity and in the wild. In addition, since it is considerably easier to study neural process in simple animals, these studies provide clues about processes that may be important in humans. Obviously, brain processes in humans and other animals differ, but studies of simpler organisms offer clues and information about human learning processes.

Research methods. Attempts to find the neural bases of learning are ultimately aimed at understanding the function of the human central nervous system. However, because many of the techniques used in investigations are inherently destructive to some degree, moral and ethical considerations preclude their use in humans.

Observation and deduction. One of the oldest and still one of the most productive techniques for the early stages of any study is the method of observation and deduction. The subject is put in a learning situation and observed for changes in behavior. Deductions can then be made about the events that take place from what is known about the function of the nervous system. I. Pavlov used that method to develop one of the most widely used learning techniques, classical conditioning, although many of Pavlov's theories about the brain substrates of learning have been shown to be incorrect.

Observation of brain-damaged humans has provided clues about brain areas that are involved in certain learning functions. While this method provides no information about the neural processes underlying various kinds of learning, the damage resulting from illness, closed head and penetrating wounds, and necessary surgical procedures have provided data on which areas of the brain, and especially the cortex, subserve many of the higher human functions. Obviously, it is ethically and morally impossible to undertake controlled lesion or ablation studies on humans, but the uncontrolled brain lesions seen in clinical medicine have been very helpful in guiding controlled studies in animals. With the advent of noninvasive scanning techniques such as positron emission tomography (PET) and computerized tomography (CT) and magnetic resonance imaging (MRI), it is now possible to provide detailed images of the human brain that reveal damaged areas in some detail. *See* MEDICAL IMAGING.

Ablation. The ablation technique involves removal of a part of the brain and observation of the effects after recovery of the subject. The underlying assumption is that removing some part of the nervous system will cause a deficit that accurately reflects the function of the lost part. This technique has been utilized successfully in brain mapping to localize various sensory and motor functions, but localization is only a partial step in characterizing the brain areas responsible for or involved in learning.

Since one brain area may be involved in more than one task, the finding that a cortical area is involved in vision does not preclude its active involvement in other tasks. However, the ablation method assumes that an area is primarily involved in one task and that the function is indeed localized there. Also, when a central brain structure is removed, any subsequent impairment in learning could be due to a loss of sensory or motor ability, a loss of communication links between areas involved in learning, or the destruction of the vital learning area.

The simplest ablations involve removing a brain area. Tissue may also be destroyed by passing an electric current through a needle into the tissue, or by suction, radiation, chemicals, freezing, and stopping blood flow to the region. Reversible ablations, which render the brain tissue temporarily inactive, can be accomplished by using chemicals (such as potassium chloride) or a cold probe.

Following ablation, the animal is allowed to recover and then is given a learning task. The results are compared with those from animals given the same task with no ablation or a different ablation. Performance deficits in the ablated animals are attributed to the loss of the lesioned area. Alternatively, the task may be learned first, the ablation performed, and then the animal tested for loss of learning. Such a design is aimed at localizing the engram or area where the effects of learning are stored, assuming that the storage and processing areas are the same.

Stimulation. The excitation of neural tissue with electric currents or chemical agents is often thought to be the opposite of ablation since one stops activity and the other starts it. However, stimulation rarely produces the opposite results from ablation and rarely has a positive effect on learning. Stimulation of an area can produce some activity that disappears if the area is ablated. The usual result of stimulation is to disrupt or slow learning, so that it can be used as a temporary lesion, to locate sites that lead to disruption of normal learning patterns when stimulated.

Electrical recording. The brain communicates through electrical signals generated by its approximately 20 billion nerve cells. Each cell can generate an electrical impulse and transmit that small electrical charge to other cells, often over substantial distances—3 ft (1 m) in the case of motor neurons that go from the spinal cord to the muscles of the feet in humans. The constant activity of many brain neurons can be recorded by electrodes either as single-cell events, as the activity of small groups of cells, or as the summed activity of millions of cells together. With large electrodes, the activity of cell populations can be recorded directly from the scalp (**Fig. 1**). *See* ELECTROENCEPHALOGRAPHY.

With the development of sensitive amplifiers, single- and multiple-cell recordings became practical. With these techniques, it is possible to implant electrodes in almost any brain area and to record activity from cells for days and often weeks with no discomfort or damage to the animal. To position the electrodes precisely, a stereotaxic instrument (**Fig. 2**) is used to hold the anesthetized animal's head securely and allow the electrode to be lowered into the brain to a precise location. Thus, animals with implanted electrodes can learn tasks while the activity in a brain area is monitored to see if the area participates in the learning. Cellular activity changes that seem to correlate with learning may be observed, but these may be unrelated to the learning of the task itself.

Developmental methods. Increasingly, learning is tested at various stages of human or animal development and the learning possible at that developmental stage is correlated with brain tissue function at the time. As with observational methods, this method gives no insight into the neural processes underlying learning, but it can give insight into the areas of the brain that subserve various forms of learning. For example, human fetuses late in the last 3 months before birth can learn movements in a Pavlovian motor learning situation through tactile stimuli administered on the mother's stomach over the uterus. Since the fetus can associate two such stimuli, this suggests that such motor learning may not require the active role of the cerebral cortex, which is not developed at that time. Likewise, studying the young of the sea snail *Aplysia*, which can learn some motor tasks, is giving investigators some insight into the parts of the nervous system underlying certain portions of the learning process.

Other methods. Increasingly powerful techniques have been under development for investigating the changes underlying learning. The injection of sensitive dyes and radioactive tracers allows delineation of changes in nerve tracts or chemical alterations. Computer analysis of neural activity or activation of neurons allows the detection of processes underlying the engram formation. Powerful microcomputers can record, stimulate, and even perform ablation studies that were once impossible or too time-consuming. In addition, computers can provide modeling of brain function. However, many of these methods provide clues about which brain areas may be involved in learning rather than the actual process of learning.

Model approaches. Investigators of the neural substrates of learning now use the strategy of operationally defining learning. This means that learning is defined by the operations used to produce it in the experimental situation. Thus, simple associative or motor learning can be defined by the operations of classical conditioning, that is, the presentation of a conditioned stimulus and an unconditioned stimulus in a specified order and time sequence. Because learning is the development of a unique response to two repeated stimuli that are presented at the same time, the behavioral change to be studied can be specified and reproduced precisely.

The use of a simple learning situation is an example of a model approach to the search for neural mechanisms of learning. The investigator can specify the learning task and simplify the complexity of the behavioral changes as well as the neural system to be studied. The human nervous system would be considered the total preparation, with human learning as the total learning system. A nonhuman species for study of a simple learning task and for analysis of the brain processes involved would be a model preparation using a model system.

One of the simplest models for studying the associative learning of Pavlovian conditioning is an invertebrate, such as *Aplysia*. *Aplysia* has no central brain but rather a system of six pairs of ganglia that control its body functions and are interconnected by nerve pathways to allow coordinated activities. In addition, many of the nerve cell bodies are very large (almost visible to the naked eye) and can be identified in each animal. *Aplysia* can learn to associate a light touch to the siphon region with an electrical shock to the tail. The shock causes the animal to quickly and fully withdraw the siphon, whereas the touch does not. Following several paired presentations of the touch (conditioned stimulus) with the shock (unconditioned stimulus), withdrawal to the touch is greatly enhanced. Because that same response does

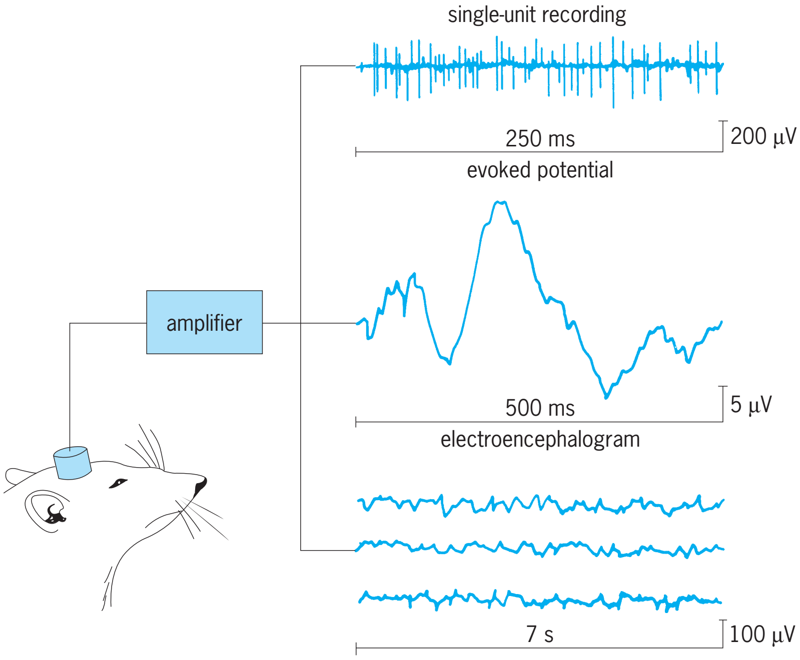


Fig. 1. **Brain activity can be recorded in various ways, from single-cell activity (top tracing) to the familiar "brain waves" or electroencephalogram (bottom tracing). (*After C. Cotman and J. McGaugh, Behavioral Neuroscience, Academic Press, 1980*)**

Fig. 2. **Stereotaxic instrument designed to allow precise placement of electrodes into brain structures. Note the four electrode carriers positioned above the area where the animal's head would be placed. (*Kopf Instrument Co.*)**

not take place in animals which receive the same stimuli in other combinations, *Aplysia* can be said to have learned the association. The neural changes underlying its behavior have been located in the abdominal ganglia and seem to involve the activation and alteration of facilitatory neurons that lie between the input sensory neurons and the output motor neurons, and a change in the output of the neurotransmitter from one neuron apparently activates the next neuron in the chain. Such findings cannot automatically be applied to vertebrates and mammals. Mammals have vastly more complex nervous systems and may utilize different neural mechanisms to achieve learning. An understanding of invertebrate learning, however, can offer clues for understanding human learning. *See* NERVOUS SYSTEM (INVERTEBRATE).

Mammalian preparations. To understand learning in the mammalian nervous system, the site at which learning occurs within that system must be found. The simplest mammalian system to be studied is the isolated spinal cord, which is capable of many reflex activities even when separated from the brain. The isolated spinal cord of the cat can alter behaviors when two stimuli are presented to selected sensory inputs in the same way as classical conditioning in the intact animal. Thus, the spinal reflex pathways apparently have the ability to "learn." The neural mechanisms of the change are obscure, but they apparently take place in the small interneurons that connect the sensory inputs with the motor outputs of the spinal cord, and may be alterations in the synaptic junctions between the interneurons. *See* REFLEX.

Classical conditioning has been shown to occur not only in the isolated spinal cord but in cats and rabbits without a cerebrum or cortex as well as in human infants born without a cerebral cortex. The ability of the mammalian nervous system to associate two meaningful stimuli and to change behavior following the association, therefore, appears to be a primitive neural attribute that is probably a basic part of all learning, including human verbal learning. Thus, associative learning appears to occur at various levels of the mammalian nervous system, and the various levels can support different complexities of learning. The total nervous system would, therefore, be necessary for the complexity of human learning. However, the question of the primary location of various learning processes remains.

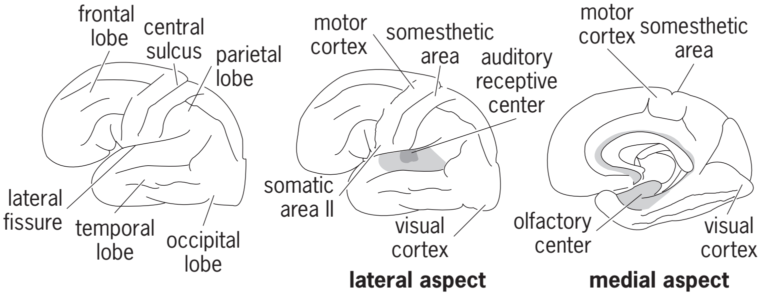


Fig. 3. Diagram of the cortical areas of the human brain. (*After S. W. Ranson and S. L. Clark, Anatomy of the Nervous System, 10th ed., Saunders, 1959*)

Perhaps the most intensive investigation of simple learning in an intact animal has utilized the eye blink response in rabbits. This model system confines the neural system to a specific sensory and motor segment of the very complex rabbit central nervous system. Since the same task is learned in the same way by humans, the learning task can be generalized to humans. By using Pavlovian conditioning, the rabbits learn to blink to a tone sounded just prior to an air puff to the eye. The hippocampus seems to produce a pattern of electrical activity that could be responsible for the learned behavior; unfortunately, the animal can learn the task equally well if the hippocampus is removed. Only when certain complicating factors are added to the simple learning task is the hippocampus found to be necessary for normal learning.

A small area in the cerebellum seems to be necessary for learning the eye blink in the rabbit. When that area is removed, the learned behavior is lost and cannot be regained. The cerebellum has long been considered a strictly motor area involved with coordination, but the learning seen in the eye blink response occurs in the interconnections of two types of neurons in the cerebellum and may involve the alteration of connections between these neurons. Thus, the primary association for learning the eye blink response occurs in the cerebellum. These studies have begun to shed light on the actual mechanisms involved in learning and storage of that learning but have also demonstrated another problem involved in such searches—locating the site of the primary change.

In more complex learning tasks, such as sensory learning, there are various levels of involvement of the higher brain structures. For example, following complete removal of the visual cortex, which subserves vision, a cat or monkey can learn to distinguish lights of different intensities and even speed of movements of visual images. Humans with damage to those visual areas of the cortex can still react to lights in the damaged area of the visual field and can even learn some tasks based on light in that area of the visual field. Some information in the visual sense is, therefore, being processed in other areas of the brain. If the cortex is not completely destroyed, some abilities to learn such things as discrimination will remain but the learning may be impaired, and often learning that had occurred prior to the lesion is lost and must be relearned. When essentially all of the cortex subserving a sense is destroyed, only rudimentary learning can occur, suggesting that some processing of the learning is done at other sites in the brain. **Figure 3** shows some of the important cortical brain areas for sensory information processing and learning in the human.

Observations of humans who have suffered brain damage or who have had brain areas removed surgically have provided information on the primary sites and mechanisms involved in mammalian learning. The most famous, "H. M.," had his hippocampus removed at an early age to control intractable epilepsy. H. M. could remember clearly everything that happened prior to surgery and had no impairment of intelligence, but after surgery he could no longer remember new facts, demonstrating that the human ability to transfer factual information into long-term memory resides in the hippocampus. However, H. M. could learn certain things. He could learn the rules of a game while he was playing it, but the next day, when asked if he remembered the game he denied ever having seen it. When asked to play, however, he played the game like anyone else who had learned it while still denying ever having seen it. Thus, the learning of the task for verbal or conscious process had not occurred, whereas the learning of motor patterns for playing the game had occurred and been stored. Thus, in humans the learning of a complex task such as a game is divided into at least two parts, one conscious and the other motor. Similar reports of human learning deficits have come from other brain-damaged humans who cannot remember learning anything following the damage but can relearn some sensory or motor tasks. The learning mechanisms are therefore intact for simpler forms of learning, but the results of that learning cannot be brought to consciousness. In addition, evidence suggests that other learning systems in the human are similarly separate. In one case, a brain-damaged accident victim could recall facts about the world and could learn new ones, but he could not remember anything about himself as a participant, even though he remembered who he was. This suggests that in the human, certain brain systems (probably in the frontal lobes of the human brain) may have evolved to support the complex learning of self-awareness and language, which is characteristic of humans.

Theories. There have been several theories formed about the actual mechanisms underlying learning process. These theories have often followed current technology, and all have been simplistic.

Electrical activity. The earliest theories proposed that learning was an electrical circuit set up in some localized area of the brain that remained active, thus explaining forgetting and other aspects of memory. However, studies of cooling the brain to a point at which there was no electrical activity or giving electrical shocks to the brain that disrupted any ongoing activity showed that the memory of learned tasks survived such treatment.

Following this, it was proposed that the initial form of learning was stored as an electrical event that was transformed over several minutes into a more permanent biochemical change in one or more brain neurons. Such notions were modeled after the growing complexity of the telephone system in the 1930s and 1940s. Learning was likened to a switchboard, with changes in individual neurons resulting from the learning activity and producing functional switchings and memories that could be recalled by the appropriate stimuli. Such "one neuron–one memory" theories had some appeal, but they failed to explain how many neurons could be lost from the brain with no apparent loss of memory for learned tasks. In addition, it became increasingly difficult to account for complex learning with the activity of one neuron.

Learning molecules. The deoxyribonucleic acid (DNA) hypotheses of the 1950s and 1960s popularized the theory of learning molecules. If learning was a change in some chemical in the brain or individual neuron, a stable chemical had to be found to account for the stability of the engram. The only sufficiently stable chemical was DNA. Further support for the DNA basis for learning came from studies purporting to show that *Planaria* flatworms that were fed other *Planaria* that had learned a simple task seemed to learn that same task by ingesting the trained flat worms. The hypothesis of learning molecules gained popularity because it suggested the feasibility of learning by pill. However, the effects of such learning molecules, if any, have been shown to be a general one due to enriched diet, and not a specific effect of transferring molecules.

Growth theories. Evidence for a similar theory has come from studies of animals raised in different laboratory environments and given various tasks to learn. Under appropriate conditions, animals in "enriched" environments, that is, with many colors, toys, and textures, have heavier brains than similar animals raised in "poor" environments. Similarly, animals given complex motor tasks to learn may develop larger brain areas associated with motor and learning function. In these studies, however, it is difficult to separate the effects of motor activity from learning itself. The growth theory of learning states that neurons involved in learning grow larger or have more neural processes when compared with noninvolved neurons. How that would occur is not certain, but using the analogy of a muscle, a neuron would enlarge with exercise and shrink with disuse.

Computer analogies. Several theories of learning have drawn analogies between the brain and computer function. These theories postulate that the brain has a central processing unit, or learning center, and that memories are stored in another location to be called up at will. In fact, the human brain does not function like a computer but rather has a vast ability to grasp general concepts and rather limited processing ability.

Activity patterns. A global theory of learning process, drawn from several earlier theories, is based on a probabilistic hypothesis in which an incoming stimulus sets up generalized patterns of activity in some brain area that is capable of causing some specific pattern of motor activity. If the incoming stimulus or a similar stimulus recurs, the same pattern of activity occurs over the same brain area and the learned behavior is executed. The theory depends on the summed performance of millions of neurons. While such a theory is elegant and accounts for many features of learning and memory, it does not specify how a stimulus would create such an activity pattern, what changes would allow that pattern to be stored or recreated by other stimuli, or how the patterns would be translated into motor output. For complex learning, the concept of neural activity pattern changes seems to best account for human learning. *See* PROBLEM SOLVING (PSYCHOLOGY). Michael M. Patterson

Bibliography. F. E. Bloom and A. Lazerson, *Brain, Mind and Behavior*, 3d ed., 2000; P. M. Groves and G. V. Rebec, *Introduction to Biological Psychology*, 4th ed., 1992; B. Kolb and I. Q. Whishaw, *Fundamentals of Human Neuropsychology*, 3d ed., 1989; C. F. Levinthol, *Introduction to Physiological Psychology*, 4th ed., 1995; R. Ornstein, L. Carstensen, and S. Patnoe, *Psychology, the Study of Human Experience*, 3d ed., 1990; E. Tulving, Remembering and knowing the past, *Amer. Sci.*, 77:361–367, July-August 1989.

Least-action principle

Like Hamilton's principle, the principle of least action is a variational statement that forms a basis from which the equations of motion of a classical dynamical system may be deduced. Consider a mechanical system described by coordinates $q_1, \ldots, q_f$ and their canonically conjugate momenta $p_1, \ldots, p_f$. The action S associated with a segment of the trajectory of the system is defined by Eq. (1), where the integral is evaluated along the given segment c of the trajectory.

$$S = \int_c \sum_j p_j dq_j \tag{1}$$

The action is of interest only when the total energy E is conserved. The principle of least action states that the trajectory of the system is that path which makes the value of S stationary relative to nearby paths between the same configurations and for which the energy has the same constant value. The principle is misnamed, as only the stationary property is required. It is a minimum principle for sufficiently short but finite segments of the trajectory. *See* HAMILTON'S EQUATIONS OF MOTION; HAMILTON'S PRINCIPLE; MINIMAL PRINCIPLES.

Assume that Eq. (2) holds, where $p_j + \delta p_j$ is canonically conjugate to $q_j + \delta q_j$.

$$S + \Delta S = \int_c \sum_j (p_j + \delta p_j) d(q_j + \delta q_j) \tag{2}$$

Neglecting second order term in Eq. (2), one obtains Eq. (3), where an integration by parts has been made, the integrated parts vanishing.

$$\begin{aligned} \Delta S &= \int_c \sum_j (p_j\, d\,\delta q_j + \delta p_j\, dq_j) \\ &= \int_c \sum_j (\delta p_j\, dq_j - \delta q_j\, dp_j) \end{aligned} \tag{3}$$

The vanishing of ΔS requires the integrand to be a perfect differential of a quantity whose end variations vanish. The coefficients of the variations $\delta q_j, \delta p_j$ need not vanish separately because the variations are not independent, the varied q's and p's necessarily being canonically conjugate, as in Eq. (4), with terms defined by Eqs. (5).

$$\sum_j (\delta p_j\, dq_j - \delta q_j\, dp_j) = dU(q, p) \tag{4}$$

$$\delta p_j = \frac{\partial U}{\partial q_j} \qquad \delta q_j = -\frac{\partial U}{\partial p_j} \tag{5}$$

Writing $U = -H\delta t$ leads to Hamilton's equations of motion, Eqs. (6). The quantity $H(q, p)$, known as the hamiltonian function, does not contain the time explicitly because $U(q, p)$ cannot be a function of the time since the end times are not fixed and in general will vary as the path is varied.

$$\dot{p}_j = -\frac{\partial H}{\partial q_j} \qquad \dot{q}_j = \frac{\partial H}{\partial p_j} \tag{6}$$

Thus, the principle is useful only for conservative systems, where H is constant.

If $H(q, p)$ consists of a part H_2 quadratic in the momenta and a part H_0 independent of the momenta, then Eq. (7) holds by Euler's theorem on homogeneous functions.

$$\begin{aligned} S &= \int_{t_1}^{t_2} \sum_j p_j \dot{q}_j\, dt \\ &= \int_{t_1}^{t_2} \sum_j p_j \frac{\partial H}{\partial p_j}\, dt \\ &= 2\int_{t_1}^{t_2} H_2\, dt \end{aligned} \tag{7}$$

Usually H_2 is the kinetic energy of the system so that the principle of least action may be written as Eq. (8), where V is the potential energy.

$$\Delta \int_{t_1}^{t_2} 2T\, dt = \Delta \int_{t_1}^{t_2} 2(E - V)\, dt = 0 \tag{8}$$

The principle of least action derives much importance from the fact that it is the action which is quantized in the quantum form of the theory. Planck's constant is the quantum of action. *See* NONRELATIVISTIC QUANTUM THEORY. Philip M. Stehle

Least-squares method

A method, developed originally by A. M. Legendre, of obtaining the best values (the ones with least error) of unknown quantities supposed to satisfy a system of linear equations of the form shown as notation (1),

$$\begin{aligned} M_{11}a_1 + M_{12}a_2 + \cdots + M_{1m}a_m &= b_1 \\ M_{21}a_1 + M_{22}a_2 + \cdots + M_{2m}a_m &= b_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots& \\ M_{n1}a_1 + M_{n2}a_2 + \cdots + M_{nm}a_m &= b_n \end{aligned} \tag{1}$$

where $n > m$. Since there are more equations than unknowns, the system is said to be overdetermined. Furthermore, the values obtained for the unknowns by solving a given selection, m in number, of the equations will differ from the values obtained by solving another selection of equations. In the physical situation, the b_i are measured quantities, the M_{ij} are known (or assumed) quantities, and the a_i are to be adjusted to their best values.

Consider a simple example. A quantity y of interest is supposed (perhaps for theoretical reasons) to be a linear function of an independent variable x. For a series of selected values $x_1, x_2, \ldots$ of x the values $y_1, y_2, \ldots$ of y are measured. The expected relation is shown as notation (2), and the problem is to find the best values of α and β that is, respectively, the slope and intercept of the line which graphically represents the function. The best values of α and β, in the least squares sense, are obtained by writing Eq. (3) and asserting that term (4) shall be minimized with respect to α and β, that is, Eqs. (5) hold. This leads to Eqs. (6) and (7), which may be solved for α and β. For m, rather than two, unknowns the generalization is obvious in principle, although the labor of solution may be great if m is large unless a high-speed electronic computer is available.

$$\begin{aligned} x_1\alpha + \beta &= y_1 \\ x_2\alpha + \beta &= y_2 \\ x_3\alpha + \beta &= y_3 \\ \cdots&\cdots \end{aligned} \tag{2}$$

$$\eta_i = y_i - (x_i\alpha + \beta) \tag{3}$$

$$\sum_{i=1}^{n} \eta_i^2 \tag{4}$$

$$\begin{aligned} \frac{\partial}{\partial\alpha} \sum_{i=1}^{n} \eta_1^2 &= 0 \\ \frac{\partial}{\partial\beta} \sum_{i=1}^{n} \eta_i^2 &= 0 \end{aligned} \tag{5}$$

$$\alpha \sum_{i=1}^{n} x_i + n\beta - \sum_{i=1}^{n} y_i = 0 \tag{6}$$

$$\alpha \sum_{i=1}^{n} x_i{}^2 + \beta \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_i = 0 \tag{7}$$

It should be noted that the measurements y_i in the example have all been assumed to be equally good. If it is known that the measurements are of variable quality, a weight may be attached to each value of y_i. The least squares equations are readily modified to take this into account, as in Eqs. (8), where w_i is the weight of measurement y_i.

$$\begin{aligned} \alpha \sum_{i=1}^{n} w_i x_i + \beta \sum_{i=1}^{n} w_i - \sum_{i=1}^{n} w_i y_i &= 0 \\ \alpha \sum_{i=1}^{n} w_i x_i{}^2 + \beta \sum_{i=1}^{n} w_i x_i - \sum_{i=1}^{n} w_i x_i y_i &= 0 \end{aligned} \tag{8}$$

The least squares equations can be shown to lead to the most probable (in the statistical sense) values of the unknowns under a variety of assumptions about the measurements and their weights. In applications in the physical sciences, however, it is rarely possible to show that one's observations satisfy all, or even any, of the assumptions. However, the conditions may be approximately satisfied in many instances, and the method is widely used because of its convenience. The empirical result is that the unknowns so determined lead to excellent representations of the data in the usual case. *See* CURVE FITTING.

McAllister H. Hull, Jr.

Bibliography. R. W. Farebrother, *Least Squares Computations*, 1988; J. W. Longley, *Least Squares Computations*, 1984.

Leather and fur processing

The technology of processing animal hides and skins. While both leather and fur are derived from animal skins, there are essential differences in the processing techniques.

Leather

Leather processing provides a means of tailoring the unique fibrous architecture of animal skins into materials that possess those attributes for which finished leather is prized: a certain texture (feel) and appearance as well as characteristic physical and chemical properties. The raw materials, primarily skins of cattle, goats, hogs, and sheep, are actually by-products of the many worldwide meat production industries. Articles manufactured from leather include footwear, clothing, gloves, luggage, upholstery, harnesses and saddlery, and mechanical devices.

Raw stock preservation. Since skins and hides putrefy rapidly after slaughter, some immediate treatment is usually necessary. While some fresh skins and hides may undergo the tannage conversion processes immediately, common salt (NaCl) is usually employed as a temporary preservative, either as a dry salt pack (**Fig. 1**) or as a saturated brine solution in an oval circulated vat (raceway) or in rotary vessels.

In the United States, hides from cattle are most frequently treated by the agitated brine curing process.

Fig. 1. Cattlehides being cured with salt in a pack containing about 800 hides.

After the hide is removed from the carcass, it is washed in cool water to lower its temperature and to remove blood and dirt. The flesh that remains on the inner side of the hide along with any manure and other debris is often removed by machine, either before or after brine curing. The brine in the raceway is maintained as close as possible to the saturation point (preferably above 90% saturation) throughout the curing process, which extends over approximately 24 h.

When properly cured, the brined hide will have a water content ranging from 40 to 48%, and the retained water will contain NaCl in excess of the 85% saturation level. Since salt-tolerant (halophilic) bacteria may contaminate recycled brines, a bactericide is added to the brine, particularly when ambient temperatures are relatively high.

Hides that have been cured with either dry salt or brine may be stored for several months, especially in refrigerated storage. This temporary stability facilitates distribution of the cured hides.

The epidermal tissue in the grain area of a typical hide (**Fig. 2**) includes the outer epidermal layers and the associated hair follicles, oil glands, sweat glands, and erector pili muscles. Collagen, a unique triple-helical biopolymer, is the leather-making hide substance; it predominates in the hide corium and is a major proteinaceous component in the grain enamel and fibrous grain layer that surrounds the epidermal tissues. Grain enamel is the extremely thin outer surface cellular layer of a skin or hide; this layer has a unique structural organization and collagen composition. The unique three-dimensional organization of the biopolymers is responsible for the strength, toughness, and flexibility of the final leather. *See* COLLAGEN; SKIN.

The finest leather is produced from skins that possess the proper grain enamel and hair or wool follicle patterns that permit the production of a unique surface appearance in the final product. Skins or hides whose grain quality is inadequate may be finished into products known as corrected-grain (a leather whose grain surface has been abraded to remove or obscure defects) or sueded leathers. Heavier hides may be mechanically split into two layers; the inner layer or split is usually converted to sueded split leather. These grade/type designations, including species of origin and grain integrity, are protected by fair trade rules.

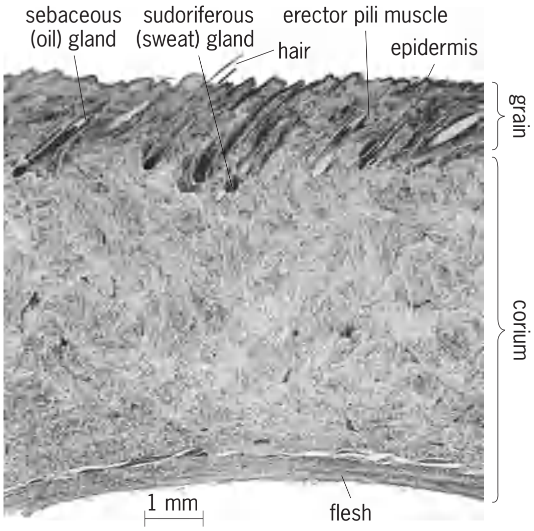


Fig. 2. Cross section through cattlehide. 1 mm = 0.04 in.

Beamhouse operations. The preserved dry skins or hides are rehydrated, and the curing salt is removed by soaking in a paddle vat, a rotary drum, or hide processor, a process requiring up to 12 h. Bactericides are added when the soaking process is longer or if warm water is used.

The rehydrated skins or hides are then processed in a saturated lime [$Ca(OH)_2$] solution with caustic sharpeners (alkaline additives), usually sodium sulfide or sodium sulfhydrate. This treatment may either loosen the hair and hair follicles for mechanical removal or pulp the keratinous epidermal tissue so that it may be washed off. Heavy hides may be processed with recovery of salable by-product hair in a 5-day vat process; this is used especially for production of sole leather. Processors containing material with lower ratios of solution to hide by weight and higher concentrations of sharpeners can pulp the hair in a few hours. However, this may require a few hours of additional treatment with lime solution to open up the hide corium to permit proper processing subsequently.

Sheepskins are dewooled by painting a thick chemical suspension for hair removal onto the flesh side. Raw material for leathers that retain the original hair or wool, including sheepskin shearlings, is washed carefully and rewetted to preserve the wool or hair in the follicle.

Chromium tannage. The beamhouse processes remove keratinous tissues from the hide and initiate the preparation of collagenous fibrous substances in the hide for tannage. Further conditioning of the grain (bating) is accomplished with enzymes; these are usually pancreatic enzymes, although sometimes bacterial or fungal enzymes are used. Deliming salts, most frequently ammonium sulfate or ammonium chloride, are employed to adjust the pH of the bath to the range of 8 to 9. *See* PH.

Chromium tannage requires an acidic pickling process. The pickling solution is a sodium chloride brine solution (using water approximately equal in weight to the hide weight and using NaCl equal to 3% of the weight of the water used) and sufficient sulfuric acid to create an acidity of pH 2–3 in the final pickling bath. Pickling, in addition to being a tannage process step, is also used for temporary storage of the skins or hides. This is an important treatment for sheepskins after wool pulling; pickled sheepskins are shipped from New Zealand to worldwide markets.

The tanning material is essentially a complex containing one-third basic chromic sulfate ($CrOHSO_4$). The uptake of the trivalent chromium by the hide substance is moderated by using so-called masking chemicals, often sodium formate. The adsorbed chromium may be fixed onto the hide collagen by pH adjustment (to approximately pH 3.8–4.0), with sodium bicarbonate being a common basification agent.

A characteristic manifestation of chromium tannage is the increase in the hydrothermal shrinkage temperature of the hide substance from around 149°F (65°C) to temperatures exceeding 212°F (100°C). The amount of the chromium tanning complex that is fixed varies, but it usually requires more than 2% fixation (expressed as the Cr ion) on the protein basis. The tanning complex is fixed by collagenous carboxyl groups, partially forming crosslinks in the collagenous molecule. The entire beamhouse and chromium tannage process for cattlehides may be accomplished in a daily cycle; however, a total time of 48 h for both processes often is preferable.

Since hides may be very greasy, degreasing agents may be utilized in the beamhouse and tannage processes. These are often aqueous solutions; however, pickled sheepskins are commonly degreased with hydrocarbon solvents.

The chromium-tanned stock is blue and is often known as chrome-in-the-blue stock. It is an article of international commerce, and is shipped in the wet stage. While the wet stock is resistant to bacterial deterioration, fungal growth can occur during storage. Hence, fungicides are commonly added to prevent damage. *See* FUNGISTAT AND FUNGICIDE.

Retannage, coloring, and fat-liquoring. The wet chrome-in-the-blue tanned hides are sorted to select the raw material most appropriate for production of the more valuable full-grain leathers. The sorting also groups the stock by hide thickness and surface area to permit the optimal production of the final grain leather thickness and types. The hides are split mechanically, producing a grain layer and a so-called blue drop. After trimming, this inner split layer is usually converted into suede leathers.

Retannage processes may include a second trivalent chromium tannage, a natural vegetable tannage, or a variety of organic synthetic tannages. These combined tannage processes permit the tanner to modify the ionic charges on the tanned collagen fibers facilitating the next two steps in this continuous trio of processes. Proper retannage procedures also facilitate the surface buffing process on the dried leather, which is required for the production of corrected-grain leathers.

Acidic, basic, direct, and reactive dyestuffs produce the colors required for the many fashion markets served by leather. The colorants are selected and applied either to produce surface color effects or to produce leather with complete cross-sectional color penetration. The levels of consumption of the dyes may vary from 1 to 10% by weight of the wet stock, particularly in suede leathers. *See* DYEING.

Fat-liquoring is a lubrication process that uses oil emulsions. It is designed to produce leathers with the suppleness, temper, and strength needed for the varied end uses. The lubrication oils are most frequently natural oils, of animal (especially marine) and vegetable origin; however, synthetic lubricants and mineral oils are also utilized to produce desired effects in the final leather. The emulsions may have either anionic, cationic, or nonionic properties that impart different characteristics to the leather. Sulfated oils, bisulfited oils, and soap emulsions are frequently used in commercial fat-liquoring systems. Levels of lubricant additives vary from a few percent to greater than 10% in oily or so-called stuffed leathers. *See* FAT AND OIL.

Retannage, coloring, and fat-liquoring processes are continuous operations carried out in rotary drums. They are accomplished in a few hours, and they are often carried out in warm water to facilitate the process. A tannery may have available several hundred different formulas for these operations in order to produce a wide enough variety of leathers to meet consumer demand.

Drying and finishing. The development of the final leather characteristics continues in the tannery processes that complete the conversion of skins or hides to leathers for varied commercial end uses. The first step is the removal of the retained process water, so that the dried leather contains about 12% moisture.

Drying processes are designed to produce leathers with the desired feel, surface levelness, and area (these leathers are sold by the square foot). Very soft glove or garment leathers may be dried by slower processes during a hanging period in a drying room with controlled air flow maintained at moderate temperatures. In toggle drying processes, the moist leather is attached to specially designed frames prior to passage through a drying tunnel; this drying process yields dry leather with the desired suppleness and with somewhat increased area yield. Firmer leathers with greater area yield may be affixed upon a drying plate coated with a temporary adhesive in a unit known as a pasted leather tunnel dryer. This process often employs higher drying temperatures and is most often used to produce leather for shoe uppers. Other methods used are vacuum drying processes, radio-frequency drying processes, and dehumidification drying processes.

The selection of the drying process is determined by the desired final leather characteristics and leather area yield. There has been considerable development of processes and machinery involving this step in the tannery process.

Leather suppleness may be further modified after drying by dry milling in a rotary drum or by mechanical flexing processes. Furthermore, dried raw material whose grain quality is inadequate is buffed with selected abrasive papers to modify the grain enamel surface prior to finishing as corrected-grain leathers.

The surface luster of the leather is developed by the application of finishing coats to the grain of the leather. These finishing formulas utilize a film-former binder, a colorant, a carrier, and additives that promote finish coat adhesion to the leather surface to achieve desired leather surface characteristics. The film-forming binder polymers may include proteins, nitrocellulose, acrylics, or polyurethanes.

The colorants are either dyes or pigments, especially on corrected-grain leathers. Water is the basic carrier of the finishing formulation, although other solvents may be used, especially in lacquer topcoats.

The finishing formulations are applied with rotary sprayers or roll coaters, with the skins or hides moving on conveyors. The carrier is removed by evaporation in heated tunnels. These finishing coat applications, which are usually repetitive, are accompanied by mechanical plating operations to facilitate the development of the desired surface luster and appearance. The finished leather is sorted by estimated cutting yield, and the surface area is measured. Packaging and shipment of the finished leather to manufacturers of leather products completes the chromium tannage conversion process.

Vegetable tannage. Vegetable materials are the initial tanning materials in approximately 10% of leather production in the United States. Vegetable tanning materials are polyphenolic extracts of barks and woods. The major commercial extracts are quebracho, wattle (mimosa), and chestnut; these materials are all imported from South America, South Africa, and the Mediterranean countries. Vegetable-tanned leathers in the United States are primarily heavy cattlehide leathers produced for shoe soles and insoles, belting, strap, and specialty leathers. *See* WOOD CHEMICALS.

Vegetable tannery processes are specialized adaptations of the general tannery processes previously discussed. The vegetable-tanned sole leather process utilizes a special hide segmentation process (**Fig. 3**). After the beamhouse operation, the bellies are pickled and used for chrome-tanned leathers for work gloves, and the heads may be used for dog bone production. The shoulders are used for specialty leathers, including waist belt and strap leathers. The bends are converted to shoe sole and insole leathers.

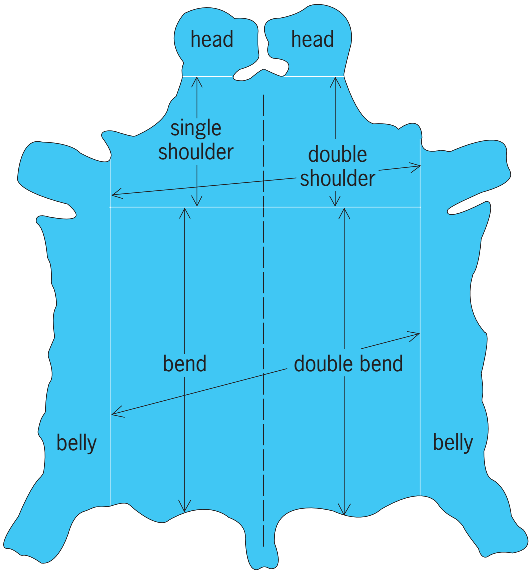


Fig. 3. Sole leather from cattlehide.

The original vegetable tannage process was a 2- to 3-week process in which the hide was sequentially immersed in a series of rocker vats containing progressively increasing tannin strength. This process resulted in incomplete tannin utilization and caused wastewater problems. A more recently developed method uses an initial polymeric phosphate and sulfuric acid pickling bath to create the acidic pH for tannage and to protect the grain of the hide. Vegetable tannage is then carried out in a concentrated tannin solution in a heated rocker-circulator system. This system provides better utilization of the vegetable tannin blend and produces the desired 50–65% combined tannin, expressed on the collagenous protein content of the leather.

The characteristics of solidity and compactness that are necessary to produce good soling leathers are developed further by using a rotary drum known as an oil wheel to apply oils and inorganic salts (for example, epsom salts). The leather is dried very slowly in hanging lofts, and then it is compacted by high-pressure rolling. The leather is not dyed; the grain surface is polished by glazing with a wax emulsion applied to the grain side of the bend.

Fur

The earliest type of clothing for primitive humans was fur, prized for warmth in cold climates and for exotic beauty, color, and texture in warmer climates. Furs possess the unique characteristics of the keratinous epidermal portion of the skin, with the collagenous corium providing the supporting fibrous matrix that enables fur to be used as a clothing material.

Raw materials. The sources of the raw materials are wild and domestic animals. Trappers provide the wild animal skins, which must be preserved under field conditions, sometimes resulting in poor quality. When animals that are bred and raised in captivity, such as mink and rabbit, are produced primarily for their fur skins, better quality control is achieved.

Animals whose skins are used for fur include beaver, fox, mink, muskrat, rabbit, sheep, and seal. Mink represents the highest-dollar-volume fur in worldwide trade. Woolskins from domestic sheep, in addition to being the source of shearlings for garments and footwear, are specially processed to produce mouton fur. The skin of the newborn lamb of the hardy fat-tailed karakul sheep is the source of a major commercial fur. Mink are bred to produce special color styles. Since rabbit skins are numerous and inexpensive, special breeds have also been developed to permit production of rabbit furs with varied colors and textures.

Fur skins possess characteristic colors and color patterns. The natural fur pigmentation is determined by the presence of melanin, an oxidized amino acid derivative that is synthesized by specialized epidermal cells.

The morphology of the epidermal system in fur skins is more complex than that found in raw materials for leather making such as calfskins. The predominant fur of the pelt contains longer hair fibers underlain by a dense, fine, soft undercoat. Some species also have coarser specialized hairs, including guard hairs.

Processing. The fur dresser rehydrates the fur skins carefully to avoid damage to the expensive raw material. Chemical additives are avoided to protect the integrity of the fur. Mechanical action is minimized, especially on fine, thinner skins. The rehydrated, softened skin is then fleshed, using special machines. Careful fleshing is a necessity, since dressing agents must penetrate from the flesh side of the corium and mechanical action must be avoided.

Tannage, or dressing, of fur skins may employ processes related to those used in leather tannage. Since garments fashioned of fur are seldom exposed to heavy rain and are not laundered, pickled skins containing salt may be merely lubricated prior to drying and finishing. However, alum or chromium tannage processes and aldehyde (formaldehyde or glutaraldehyde) tannages may be employed to stabilize the corium protein further, providing increased resistance to water and heat. Excess fixation of chromium onto the fur protein can lead to color variations in the fur. Chromium tannages may also interact with bleaching processes and dyeing procedures involving oxidation.

Oiling of fur skins is accomplished by applying light oils with low viscosity to the flesh side of the skin. Fat-liquoring operations using oil emulsions in rotary drums would produce fur entanglement and wool felting. Heavier skins may have the oil worked into the skin corium in a device known as a kicker tub; unsaturated oils may be used in this process for lubrication and to produce an oxidized oil tannage. Adjustment of the fur color to produce fashionable colors is a complex process; its use is regulated to assure correctness of labeling.

In fur processing, the hydrothermal stabilization of a skin substance is often low; thus removal of excess water requires care. Centrifugal processes may be used, but the corium retains moisture more tenaciously than the fur fibers. Hanging the fur skins to dry under controlled conditions of temperature and humidity in order to avoid damage completes the process. However, if the corium has been dried to a 15% moisture content, the fur will often be overdried, with low flexibility and excess static electric charge. Agitation in a drum containing sawdust is used to equalize moisture content in the corium and fur and to dryclean the fur. After sawdust removal and adjustment of the corium moisture to about 25%, careful stretching or staking develops skin softness.

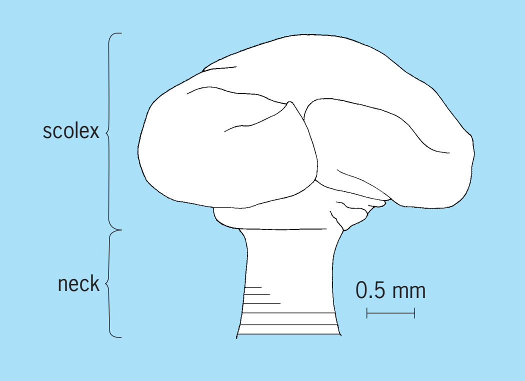
Selective removal of guard hairs and shearing are special mechanical processes that may be used. The fur gloss is developed by ironing in rotary ironing machines.

Robert M. Lollar

Bibliography. K. Bienkiewicz, *Physical Chemistry of Leather Making*, 1983; F. O'Flaherty, W. T. Roddy, and R. M. Lollar, *The Chemistry and Technology of Leather Manufacturing*, Amer. Chem. Soc. Monogr. 134, 4 vols., 1956–1965, reprint 1978; G. C. Moseley, *Leather Goods Manufacture: Methods and Processes*, 1986; J. H. Sharphouse, *Leather Technician's Handbook*, 1983; T.C. Thorstensen, *Practical Leather Technology*, 4th ed., 1993.

Lecanicephaloidea

An order of tapeworms of the subclass Cestoda. All species are intestinal parasites of elasmobranch fishes. These tapeworms are distinguished by having a peculiar scolex divided into two portions. The lower portion is collarlike and bears four small suckers; the upper portion may be discoid or tentacle-bearing and is provided with glandular structures (see **illus.**). The scolex is usually buried in the intestinal wall of the host and may produce local pathology. The anatomy of the segments is very similar to that of the Proteocephaloidea. Some authorities place the lecanicephaloids in the order Tetraphyllidea, to which they are obviously closely related. Essentially nothing is known of the life history of lecanicephaloids. *See* EUCESTODA; TETRAPHYLLIDEA.



Anterior end of a lecanicephaloid tapeworm.

Clark P. Read

Lecanorales

An order of the Ascolichenes, also known as the Discolichenes. Lecanorales is the largest and most typical order of lichens and parallels closely the fungal order Helotiales. The apothecia are open and discoid, with a typical hymenium and hypothecium. There are four growth forms—crustose, squamulose, foliose, and fruticose—all showing greater variability than any other order of lichens.

Reproduction. Details of sexual reproduction are at best poorly known. Meiotic and mitotic stages leading to spore formation seem normal, but earlier stages of fertilization have not been studied. It is not known if fertilization is somatogamous, or if the microconidia act as spermatia and fertilize the egg in the ascogonium. Mature spores burst the ascal wall and are ejaculated from the apothecial disk to heights of several centimeters. When the spores are germinated in pure culture, a distinct though slow-growing mycelium is often formed. The mycelium has no resemblance to the growth form of the original parent lichen, nor have apothecia or pycnidia ever been seen in culture. Furthermore, it is doubtful whether anyone has ever synthesized a lichen by combining a fungal culture with an alga. In nature vegetative reproduction is highly developed. Isidia, minute coralloid outgrowths of the cortex, and soredia, powdery excrescences from the medulla, are present in many species. They apparently grow into new lichens when dislodged from the thallus. *See* REPRODUCTION (PLANT).

Taxonomy. The Lecanorales is divided into 25 families, about 160 genera, and 8000–10,000 species. Family divisions are based on growth form of the thallus, structure of the apothecia, the species of symbiotic algae present, and spore characters. Species are separated by such characters as isidia, soredia, rhizines, and pores, and by chemistry. The larger families are described below.

Cladoniaceae. This family includes the reindeer mosses and cup lichens. The main thallus is a hollow structure in the shape of a stalk, cup, or is richly branched. There are about 200 species of *Cladonia*; the most highly developed are found in the boreal and arctic regions.

Lecanoraceae. The thallus is typically crustose, and the apothecia have a distinct thalloid rim. This family is primarily temperate and boreal. The largest genus, *Lecanora*, has over 500 species.

Lecideaceae. This family differs from the Lecanoraceae in lacking a thalloid rim around the apothecia. It is also very common in temperate and boreal zones and includes one of the largest and most difficult lichen genera to diagnose, *Lecidea*, with over 1000 species.

Parmeliaceae. This family, known as the foliose shield lichens, is common on trees throughout the world. *Parmelia* has almost 1000 species; *Cetraria*, a common genus on pine trees, includes the well-known Iceland moss (*Cetraria islandica*).

Umbilicariaceae. The thallus is circular in outline, quite large, and umbilicate. The family occurs on acidic rocks in temperate, boreal, and arctic regions. The members are known as the rock tripes.

Usneaceae. The beard lichens comprise perhaps the outstanding lichen family because of their conspicuous fruticose growth form. The largest genus is *Usnea*, some species of which grow to 5 ft (1.5 m) in length.

Mason E. Hale

Lectins

A class of proteins of nonimmune origin that bind carbohydrates reversibly and noncovalently without inducing any change in the carbohydrate. Since their discovery in 1888, lectins have been labeled agglutinins, antibodylike molecules, hemagglutinins, heteroagglutinins, heterophile agglutinins, natural or normal antibodies, protectins, phytohemagglutinins, and receptor-specific proteins, among other names. The main feature of lectins is their ability to recognize and bind specific carbohydrate structures becoming the translators of the sugar code (the coding of biological information by sugar structures). Lectins bind a variety of cells having cell-surface glycoproteins (carbohydrate-bound proteins) or glycolipids (carbohydrate-bound lipids). The presence of two or more binding sites for each lectin molecule allows the agglutination (clumping) of many cell types, and the agglutination reaction has been used extensively to detect the presence of lectins in extracts from different organisms. However, multivalency (having several sites of attachment) may not be an absolute requirement, even though it is still an important factor for most lectins. Lectins are ubiquitous in nature and may have very different roles according to the organism, tissue, or developmental stage. Their potential ligands, simple or complex carbohydrates, are present in all living cells and in biological fluids, which suggests that protein–carbohydrate interactions constitute basic phenomena common to all organisms. *See* CARBOHYDRATE; PROTEIN.

Distribution. Lectins have been identified in species of virtually all taxa, from viruses and bacteria to vertebrates, and they occur in both prokaryotic and eukaryotic organisms.

Agglutinating activity has been found in more than 1000 plant taxa. At present, around 1000 sequences have been determined for 241 different plants. Moreover, most of the best-characterized lectins come from a single family, Leguminosae, and they constitute about 60% of the well-known lectins. Lectins have been found in almost every plant tissue, but their distribution in different organs varies within the various families already studied. They are abundant in seeds (Leguminosae, Euphorbiaceae, Gramineae), fruits, tubers and bulbs (Solanaceae, Liliaceae), roots (Cucurbitaceae, Leguminosae), and stems and leaves (Cactaceae, Orchidaceae). At the subcellular level, they are present in plant cell walls of almost all tissues, and may have a fundamental role in the assembly of cell-wall polysaccharides and glycoproteins. *See* PLANT METABOLISM.

Recent years have seen an explosion of interest and research in animal lectins. Good examples are galectins (galactose-binding lectins); calcium-dependent (C-type) lectins, which include selectins (membrane-bound lectins specific for leukocyte adhesion) and collectins (mannose-specific lectins); and annexin proteins with affinity for lipids and sugars present in almost all organisms. Vertebrate lectins occur as soluble or integral membrane proteins in embryonic and adult fluids, organs, and tissues. Invertebrate lectins occur in body fluids or secretions, such as fish serum, snake venom, seminal and coelomic fluids, and hemolymph.

Microbial lectins have been isolated mainly from bacteria, but they are also found in viruses, slime molds, protozoa, green algae, and fungi. In many cases the microbial lectins are bound to the bacterial surface and have only one recognition site. The bacteria have the ability to agglutinate other cells, but the isolated lectin does not and is referred to as lectinlike protein.

Initially, lectins were classified according to their saccharide specificity. However, the introduction of the techniques of molecular biology has changed lectin classification. The homology in the structure of the carbohydrate recognition domain allowing for the drawing of genealogical trees has replaced this criterion. Lectins are classified based on their amino acid sequence homology and relatedness. The crystal structures of more than 500 lectins of different origin have been determined, and a very complete database is available with the known three-dimensional structures.

Molecular properties. The molecular structures of lectins from different organisms show no general pattern. This is not surprising if one considers that plants and animals have undergone long evolutionary divergence. Carbohydrate-binding molecules appear to have arisen at different times and are products of very different evolutionary histories.

Lectins can be monomeric or oligomeric. The latter are usually composed of two to four polypeptide chains and have one binding site per chain. However, they are very different in their chemical and physical properties. Most have covalently linked carbohydrates, which account for 50% in the potato lectin and only 0.5% in the lectin of the sponge *Axinella polypoides*. Some, like the jack bean lectin, concanavalin A, are completely devoid of sugars. The molecular weights of the native oligomers range from 8500 for the lectin of the stinging nettle (*Urtica dioica*) to 600,000 for the lectin from the tunicate *Halocynthia roretzi*. On the other hand, lectins from species that are closely related taxonomically exhibit high homology of the gene sequences encoding for those proteins, as has been shown for animal lectins and legume or cereal lectins.

Some lectins require calcium, as in the animal C-type lectins; others require divalent cations (calcium, magnesium, or manganese) to bind their carbohydrate ligands, as in legume lectins. Other lectins do not require divalent cations for binding activity and instead call for the presence of thiol groups, as in the wheat germ agglutinin or the soluble vertebrate lectins (galectins). The sugar specificity of a lectin is defined by the carbohydrate for which it shows the highest affinity, which means that lectins considered specific for one monosaccharide may also bind other structurally related carbohydrates having a lower affinity. In general, lectins interact with the nonreducing glycosyl groups of polysaccharides and glycoproteins, but some can bind internal sugars or sugars at the reducing end. Some lectins can recognize one particular monosaccharide, which suggests the presence of a small binding site, whereas others bind preferentially to trisaccharides or tetrasaccharides and have an extended binding site. Lectins binding oligosaccharides are thought to be more specific for their ligand than those binding monosaccharides.

Another important property of some lectins is their ability to induce cell proliferation of particular tissues or cells. Such mitogenic lectins are used extensively in biomedical research.

It seems clear that a wide variety of proteins share the ability of lectins to bind carbohydrates. Lectins from closely related organisms probably stem from common ancestors and have high degrees of homologies. That is particularly clear in the relatively less complex structure of plant lectins. However, lectins related to a specific function may or may not have structural homologies, depending on their evolutionary history. In addition, the more complex structure of vertebrate lectins indicates that they are built with different polypeptide domains that present homologies to many unrelated proteins. The carbohydrate recognition domain is the only one offering a certain degree of homology.

Biological functions. From the known examples, it seems clear that lectins are involved in recognition phenomena, and their ability to bind particular carbohydrate structures is the key to their biological functions. These recognition functions include their involvement in interactions with cells or extracellular materials from the same organism (which could be considered self-recognition, or recognition of endogenous ligands) and interactions with foreign particles or cells (recognition of non-self, or recognition of exogenous ligands). *See* IMMUNOLOGICAL PHYLOGENY.

Animal lectins are involved in a variety of cellular processes, including intracellular transport and extracellular assembly, cell–cell and cell–matrix interactions, cell growth control and apoptosis (programmed cell death), and immune functions. Microbial lectins largely function in host cell attachment, tissue colonization, and invasion. Plant lectins have been known for more than 100 years. Although their biological roles are not fully understood, the plant lectins are involved in deposition of storage proteins, maintenance of seed dormancy, defense against pathogen and animal predators, symbiosis, transport of carbohydrates, mitogenic stimulation of embryonic plant cells, assembly and elongation of cell walls, and recognition of pollen.

Lectins as tools. Lectins are very useful reagents for the study of complex carbohydrates and cell surfaces, for the separation and identification of particular cells, and for the stimulation of cell proliferation. Lectins are the main tools for glycan (polysaccharide) profiling aiming at comprehensive elucidation of glycan functions in all organisms. Lectins covalently attached to insoluble matrices are used to separate glycoproteins or glycopeptides that contain different carbohydrates. Labeled lectins are also used in histochemical and cytochemical studies to localize glycoconjugates that carry particular sugars. This technique is particularly interesting, since changes in lectin-binding patterns occur during embryonic differentiation, malignant transformation, aging, and some pathological conditions. The spectrum of uses for lectins as specific reagents continues to expand.

Rafael Pont-Lezica

Bibliography. D. C. Kilpatrick, *Handbook of Animal Lectins: Properties and Biomedical Applications*, Wiley, 2000; N. Sharon and H. Lis, *Lectins*, 2d ed., Kluwer Academic, 2003.

Lecythidales

An order of flowering plants, division Magnoliophyta (Angiospermae), in the subclass Dilleniidae of the class Magnoliopsida (dicotyledons). The order consists of the single family Lecythidaceae, with about 400 species. They are tropical, woody plants with alternate, entire leaves, valvate sepals, separate petals, numerous centrifugal stamens, and a syncarpous, inferior ovary with axile placentation. Brazil nuts are the seeds of *Bertholletia excelsa*, a member of the Lecythidaceae. *See* BRAZIL NUT; DILLENIIDAE; MAGNOLIOPHYTA; PLANT KINGDOM. Arthur Cronquist

Legendre functions

Solutions to the differential equation $(1 - x^2)y'' - 2xy' + v(v + 1)y = 0$.

Legendre polynomials. The most elementary of the Legendre functions, the Legendre polynomial $P_n(x)$ can be defined by the generating function in Eq. (1). More explicit representations are Eq. (2), and the hypergeometric function, Eq. (3).

$$(1 - 2xr + r^2)^{-1/2} = \sum_{n=0}^{\infty} P_n(x) r^n \qquad (1)$$

$$P_n(x) = \frac{(-1)^n}{2^n n!} \frac{d^n}{dx^n}(1 - x^2)^n \qquad (2)$$

$$P_n(x) = {}_2F_1[-n, n + 1; 1; (1 - x)/2] \qquad (3)$$

See HYPERGEOMETRIC FUNCTIONS.

Generating function (1) implies Eq. (4).

$$(a^2 - 2ar\cos\theta + r^2)^{-1/2} = \frac{1}{a}\sum_{n=0}^{\infty} P_n(\cos\theta)\,(r/a)^n \qquad 0 < r < a \qquad (4)$$

The function $(a^2 - 2ar\cos\theta + r^2)^{-1/2}$ represents the potential in an inverse square field at a point P of a source at A, where r and a are the distances from P and A to a fixed point O, and θ is the angle between the segments PO and OA. These functions were extensively studied by A. M. Legendre and P. S. Laplace because they could be used in the study of the celestial mechanics. They had arisen slightly earlier in probabilistic work of J. Lagrange. *See* POTENTIALS.

Since Legendre polynomials are the zonal spherical harmonics on the surface of the unit sphere in three-space, they arise when studying physical phenomena associated with spherical geometry. They arise in many other applications. One application is to the estimation of the smallest eigenvalue of the truncated Hilbert matrix $(a_{i,j})_0{}^N$, $a_{ij} = (i + j + 1)^{-1}$. This matrix is very hard to invert numerically because the condition number, or the ratio of the largest to the smallest eigenvalue, is very large. One of the essential parts of the proof is an asymptotic formula for Legendre polynomials of the interval $[-1,1]$. *See* MATRIX THEORY; SPHERICAL HARMONICS.

Relation to trigonometric functions. When $x = \cos\theta$, Legendre polynomials have a relation with trigonometric functions, given by Eq. (5), where $|R(n,\theta)| \leqq A/(n \sin\theta)$, A being a fixed constant, when $c/n \leqq \theta \leqq \pi - c/n$ for $c > 0$. *See* TRIGONOMETRY.

$$(\sin\theta)^{1/2}\left(n + \frac{1}{2}\right)^{1/2} P_n(\cos\theta) = \sqrt{\frac{2}{\pi}}\cos\left[\left(n + \frac{1}{2}\right)\theta - \frac{\pi}{4}\right] + R(n,\theta) \qquad (5)$$

Properties. A graph of $P_n(x)$ [**Fig. 1**] illustrates a number of properties. $P_n(x)$ is even or odd as n is even or odd, that is, $P_n(-x) = (-1)^n P_n(x)$. All the zeros of $P_n(x)$ are real and lie between -1 and 1. The zeros of $P_{n+1}(x)$ separate the zeros of $P_n(x)$. $P_n(x)$ satisfies the inequality $|P_n(x)| \leqq 1$, $-1 \leqq x \leqq 1$. The successive relative maxima of $|P_n(x)|$ increase in size as x increases over the interval $0 \leqq x \leqq 1$. The closest minimum to $x = 1$ of $P_n(x)$ called $\mu_{1,n}$, satisfies $\mu_{1,n} < \mu_{1,n+1}$. Similar inequalities hold for the first maxima of $P_n(x)$ to the left of $x = 1$, and for the kth relative maxima or minima. All of these results have been used in applications. The last two results about the relative maxima were used to obtain bounds on the phase of scattering amplitudes. From an important limiting relation of F. G. Mehler, Eq. (6), it is easy to show that $\mu_{1,n}$ approaches the minimum value of the Bessel function J_0. *See* BESSEL FUNCTIONS.

$$\lim_{n\to\infty} P_n\left(\cos\frac{\theta}{n}\right) = J_0(\theta) \qquad (6)$$

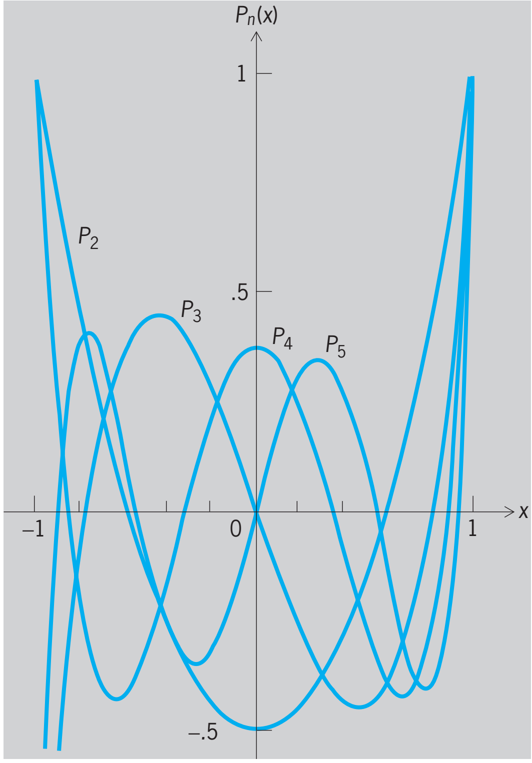


Fig. 1. Legendre polynomials $P_n(x)$, $n = 2, 3, 4, 5$.

Differential equation. One of the main reasons for the occurrence of Legendre polynomials is a differential equation they satisfy, Eq. (7). This equation arises

$$(1-x^2)y'' - 2xy' + n(n+1)y = 0 \qquad (7)$$

$$y = P_n(x)$$

from the solution of Laplace's equation by separation of variables in spherical coordinates. There is a second solution $Q_n(x)$ given by Eq. (8), when x is not in the interval $-1 < x < 1$. For $-1 < x < 1$, $Q_n(x)$ is defined by Eq. (9). $Q_n(x)$ had many of the same properties as $P_n(x)$, illustrated by **Fig. 2**. *See* DIFFERENTIAL EQUATION; LAPLACE'S DIFFERENTIAL EQUATION.

$$Q_n(x) = \frac{1}{2}\int_{-1}^{1} \frac{P_n(t)}{x-t}\,dt \qquad (8)$$

$$Q_n(x) = \lim_{\epsilon\to 0^+} \frac{Q_n(x+i\epsilon)+Q_n(x-i\epsilon)}{2} \qquad (9)$$

Addition formula. One of the most important formulas satisfied by Legendre polynomials is the addition formula, Eq. (10). The functions $P_n^m(x)$ are defined by Eq. (11). These functions, often called associated Legendre functions, satisfy differential equation (12).

$$\begin{aligned} &P_n(\cos\theta\cos\varphi + \sin\theta\sin\varphi\cos\chi) \\ &\quad = P_n(\cos\theta)P_n(\cos\varphi) + 2\sum_{m=1}^{n}\frac{(n-m)!}{(n+m)!} \\ &\qquad \cdot P_n^m(\cos\theta)P_n^m(\cos\varphi)\cos m\psi \qquad (10) \end{aligned}$$

$$P_n^m(x) = (-1)^m(1-x^2)^{m/2}\frac{d^m}{dx^m}P_n(x) \qquad (11)$$

$$-1 < x < 1 \quad m = 0, 1, \ldots, n$$

$$\begin{aligned} &(1-x^2)\frac{d^2y}{dx^2} - 2x\frac{dy}{dx} \\ &\quad + \left[n(n+1) - \frac{m^2}{1-x^2}\right]y = 0 \qquad (12) \end{aligned}$$

General Legendre functions. In work of F. G. Mehler on electrical distribution with conical symmetry, solutions arose for Eq. (12) with $m = 0$ and $n = -{}^1\!/_2 + it$, where t is a parameter that can assume any real value. When Laplace's equation is solved by separation of variables in toroidal coordinates, this equation occurs with $n = \nu - {}^1\!/_2$ and $m = \mu$, where μ and ν are separation parameters. Thus there are reasons for considering solutions to Eq. (12) when n and m are arbitrary complex numbers. In the general context, n and m are usually called ν and μ. $P_\nu^\mu(z)$ is defined by Eq. (13) for $|z-1| < 2$, with the cut $z < 1$ removed. By analytic continuation, this function is extended to the plane cut from 1 to $-\infty$ and then defined for $-1 < z < 1$ by Eq. (14). There is a second solution to differential equation (12) which is called $Q_\nu^\mu(z)$. The easiest way to carry out the analytic extension of Eq. (13) and to define $Q_\nu^\mu(z)$ is to use the general theory of hypergeometric functions. Legendre functions are just appropriate algebraic functions times a hypergeometric function of the type that has a quadratic transformation. The theory of Legendre functions up to but not including the addition formula can be developed very easily in this fashion.

$$\begin{aligned} P_\nu^\mu(z) &= \frac{1}{\Gamma(1-\mu)}\left(\frac{z+1}{z-1}\right)^{\mu/2} \\ &\quad \cdot {}_2F_1[-\nu, \nu+1; 1-\mu; (1-z)/2] \qquad (13) \end{aligned}$$

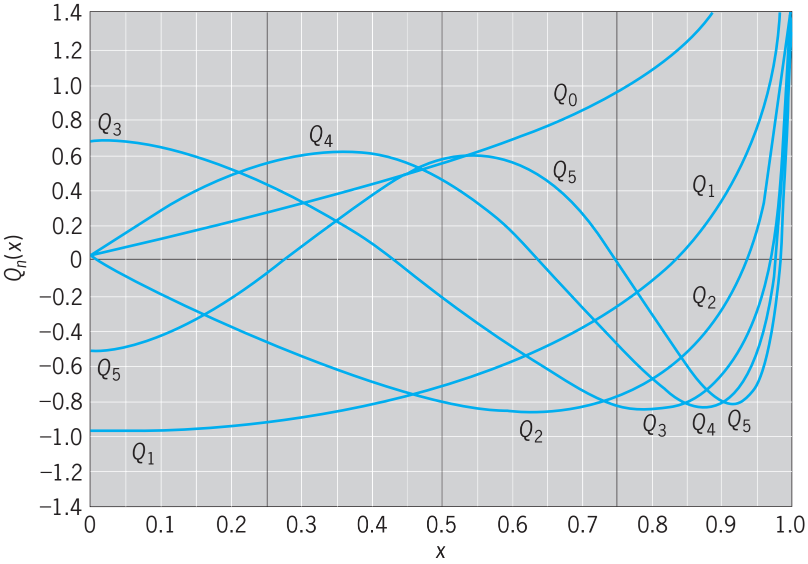


Fig. 2. Legendre functions of the second kind, $Q_n(x)$. (*After E. Jahnke and F. Emde, Tables of Functions with Formulae and Curves, 4th ed., Dover Publications, 1945*)

$$P_\nu^\mu(x) = \lim_{\epsilon\to 0^+}\frac{P_\nu^\mu(x+i\epsilon)+P_\nu^\mu(x-i\epsilon)}{2} \qquad (14)$$

Addition formulas are best derived by interpreting the functions as spherical harmonics and using the rotation groups which operate on spheres and then analytically continuing the resulting formulas in the appropriate parameters. Addition formula (10) can be extended not only to the general Legendre function of the first kind, $P_\nu^\mu(z)$, but to the general function of the second kind $Q_\nu^\mu(z)$. A number of results can be obtained from this, such as Eq. (15). This is an extension of the formula $\cos^2 x + \sin^2 x = 1$ to Legendre functions.

$$\begin{aligned} &[P_n(x)]^2 + \frac{4}{\pi}[Q_n(x)]^2 \\ &\quad = \frac{4}{\pi^2}\int_1^\infty Q_n[x^2 + (1-x^2)z](z^2-1)^{-1/2}\,dz \qquad (15) \end{aligned}$$

$$-1 < x < 1$$

Gegenbauer functions. There is a related set of functions, called Gegenbauer and associated Gegenbauer functions, which arise as spherical harmonics on higher-dimensional spheres. They are really equivalent to general Legendre functions, since they are algebraic functions times hypergeometric functions which have a quadratic transformation. The only difference is that different algebraic functions are used as multipliers since the functions satisfy different second-order differential equations. A completely analogous theory has been developed for them. There are two extreme cases of Gegenbauer functions which are very important, and in many respects their theory is easier to develop and is used as a model to suggest further developments. One is connected with the unit sphere in two-space (or the circle $x^2 + y^2 = 1$), and the resulting functions are $\cos \lambda\theta$ and $\sin \lambda\theta$. The other comes from an "infinite dimensional sphere" (actually an appropriate limit of finite dimensional spheres), and the spherical functions are Hermite polynomials and their extensions to Hermite functions. *See* ORTHOGONAL POLYNOMIALS.

Richard Askey

Bibliography. A. Erdéelyi et al., *Higher Transcendental Functions*, vols. 1–3, 1953–1955, reprint 1981; E. W. Hobson, *The Theory of Spherical and Ellipsoidal Harmonics*, 1931, reprint 1975; C. Müller, *Spherical Harmonics*, 1966; I. N. Sneddon, *Special Functions of Mathematical Physics and Chemistry*, 3d ed., 1980; N. Ja. Vilenkin, *Special Functions and the Theory of Group Representations*, 2d ed., 1983, reprint 1988; Z. X. Wang and D. R. Gue, *Special Functions*, 1989.

Legionnaires' disease

A type of pneumonia usually caused by infection with the bacterium *Legionella pneumophila*, but occasionally with a related species (such as *L. micdadei* or *L. dumoffii*). More than 40 species of *Legionella* are known, though not all have been associated with disease. The disease was first observed in an epidemic among those attending an American Legion convention in Philadelphia, Pennsylvania, in 1976. The initial symptoms are headache, fever, muscle aches, and a generalized feeling of discomfort. The fever rises rapidly, reaching 102–105°F (32–41°C), and is usually accompanied by cough, shortness of breath, and chest pains. Abdominal pain and diarrhea are often present. Diagnosis is often difficult, though there are several rapid diagnostic kits, including a urine antigen test that can augment chest x-ray presentation of a patchy interstitial infiltrate (atypical pneumonia in the case of Legionnaires' disease). Proper diagnosis is important since the mortality rate can be as high as 15% in untreated or improperly diagnosed cases. Erythromycin, new-generation fluroquinolones, and rifampicin are considered highly effective medications, whereas the penicillins and cephalosporins are ineffective.

While epidemics of Legionnaires' disease (also referred to as legionellosis) can often be traced to a common source (cooling tower, potable water, or hot tub), most cases seem to occur sporadically. It is estimated that *Legionella* spp. account for approximately 4% of all community- and hospital-acquired pneumonia. Legionnaires' disease is most fequently associated with persons of impaired immune status. Risk factors include smoking, advanced age, diabetes, underlying illness such as cancer, and immunosuppression associated with transplants, in particular, heart transplants. *Legionella* bacteria are commonly found in fresh water and moist soils worldwide and are often spread to humans through inhalation of aerosols containing the bacteria. The legionellae are facultative intracellular parasites of fresh-water protozoa, including most species of amebas. Amebas are also considered as biological amplifiers of *Legionella* bacteria and may be associated with aerosol transmission. Legionnaires' disease is not a communicable disease (no person-to-person spread as occurs with the flu), indicating that human infection is not part of the survival strategy of these bacteria. Therefore, the legionellae are considered opportunistic pathogens of humans. It is technology (air conditioning) and the ability to extend life through medical advances (such as transplantation and treatments for terminal diseases) that have brought these bacteria into proximity with a susceptible population.

Human alveolar macrophages are the primary target cell colonized by *L. pneumophila*, in which the bacteria block early events associated with the fusion of secondary lysosomes with bacteria-laden phagosomes. Instead, the phagosomes associate with and become enveloped by the endoplasmic reticulum, forming a specialized replicative phagosome, permissive for bacterial multiplication. Since this scenario occurs in natural amebic hosts, it is generally believed that the bacteria employ a common infection strategy. The mechanisms by which these bacteria alter host cell biological functions are unknown, and therefore *Legionella* infections of natural and mammalian hosts serve as model systems for study of pathogenesis and of eukaryotic cell biology.

For most humans exposed to *L. pneumophila*, infection is asymptomatic or short-lived. This is attributed to a potent cellular immune response in healthy individuals that includes gamma-interferon production by T lymphocytes as well as quick action by natural killer cells and cytotoxic T cells that kill infected macrophages. Gamma-interferon-activated macrophages restrict the intracellular replication of *L. pneumophila*. In immunocompromised individuals, a lack of gamma-interferon production (suppressed T cells) permits rapid multiplication of the bacteria in alveolar macrophages and development of pneumonia. Recovery from Legionnaires' disease often affords immunity against future infection. However, no vaccine exists at the present time. Most species of *Legionella* can be routinely grown in the laboratory on a special yeast extract medium containing activated charcoal, supplemented with the amino acid cysteine and a source of ferric iron. Bacterial colonies often take 4–7 days to appear, and the bacteria are gram-negative aerobic rods that sometimes display motility. The bacteria can survive in water for long periods of time following development of a cystlike resting form. *See* MEDICAL BACTERIOLOGY; PNEUMONIA.

Paul S. Hoffman

Bibliography. B. S. Fields, The molecular ecology of legionellae, *Trends Microbiol.*, 4:286–290, 1996; P. S. Hoffman, Invasion of eukaryotic cells by *Legionella pneumophila*: A common strategy for all hosts, *Can. J. Infect. Dis.*, 8:139–146, 1997; W. C. Winn, Legionnaires disease: Historical perspective, *Clin. Microbiol. Rev.*, 1:60–81, 1988.

Legume

A member of the plant family Leguminosae, and the name for the fruit produced by members of that family, as idealized in a bean or pea pod. The legume and grass families are by far the world's most important sources of food. Whereas grasses supply starch, legumes, which include beans, soybeans, peas, and alfalfa, supply protein and fats. Through their root nodules, which are inhabited by *Rhizobium* bacteria, the legumes also preserve the nitrogen balance in the soil. *See* FABALES; GRASS CROPS; NITROGEN FIXATION.

Legumes comprise one of the largest plant families in the world, with perhaps 18,000 species. Most species are tropical and include trees, woody vines, and herbaceous plants. All legume plants bear the same type of fruit, the pod or legume, which is diversely modified within the family. Most have compound leaves, and many, particularly tropical legumes, have spectacular flowers, some of which are grown as ornamentals.

The three major groups of Leguminosae differ in the appearance of their flowers, and some botanists consider them to be three separate families. Others, as here, regard them as a single diverse family because all have the same kind of fruit and because the groups intergrade. The first group is the Mimosoids, in which the flowers have reduced petals and usually have conspicuous, long stamens that give color to the flower clusters. The Mimosoid group, which is almost entirely tropical, includes the so-called sensitive plant and the silk tree, or mimosa, a widely planted ornamental in the southern United States. The second group, the Caesalpinioids, is characterized by well-developed petals and includes ornamentals such as the redbud and honey locust in temperate regions and the orchid trees of the tropics. The third and largest group is the Papilionoids, whose members bear a flower that resembles the sweet pea, with a big petal at the top, a wing on either side, and a "keel" of two fused petals that enclose the stamens at the bottom. Many important agricultural plants are members of the Papilionoid group, including peanuts, garden beans, soybeans, garden peas, lentils, chickpeas, cowpeas, clovers, and alfalfa. *Leuceana*, *Prosopis*, and other fast-growing woody genera provide firewood, animal forage, nitrogen fixation, and timber in the tropics. Other species yield natural fibers, flavorings, and medicinal compounds. Many are planted as ornamentals, especially in warm regions and the tropics. At least a dozen tropical species growing in the wild or primitively cultivated, especially in Africa, are locally important for human consumption. *See* ALFALFA; BEAN; CLOVER; COWPEA; LENTIL; PEA; PEANUT; SOYBEAN.

Many economic legumes have not yet been studied scientifically with regard to gene pool exploitation. Considering that presently cultivated wheat and soybeans are far different from the wild plants from which they were derived, these undeveloped legume species may well be a resource for the future. *See* LEGUME FORAGES; NITROGEN-FIXING TREES.

Duane Isely

Bibliography. Y. P. Bajaj (ed.), *Legumes and Oilseed Crops I*, 1990; D. Isely, Leguminosae and *Homo sapiens*, *Econ. Bot.*, 36:46–70, 1982; R. H. Mathews, *Legumes: Chemistry, Technology, and Human Nutrition*, 1989; J. M. Vincent, *Nitrogen Fixation in Legumes*, 1982.

Legume forages

Plants of the legume family used for livestock feed, grazing, hay, or silage. Legume forages are usually richer in protein, calcium, and phosphorus than other kinds of forages, such as grass. The production, preservation, and use of forage legumes require special skills on most soils. One important requirement is a supply of the needed symbiotic nitrogen-fixing bacteria if these are not already in the soil commercial cultures for various strains of these bacteria can be purchased and applied to the legume seed just before planting. Additional lime and commercial fertilizers may be needed on all except fertile soils. Protection from weeds, injurious insects, diseases, and other harmful influences is often required. *See* FERTILIZER; LEGUME; NITROGEN FIXATION.

Legume forages may be preserved for future use as dry hay or silage. To obtain high quality, most legume hay crops must be harvested before the mature stage and thoroughly cured without the loss of leaves; high-moisture hay spoils in warm weather. Legumes are more difficult to preserve as silage than grasses due to their high protein content. However, the addition of a high-carbohydrate material, such as molasses, or, as is usually preferred, the mixing of grass and legume are helpful measures. Legume silage often has an objectionable odor if fermentation is not sufficiently rapid to prevent breakdown of proteins.

Much of the legume forage is grazed. Some crops can be grazed continuously without injury while others require intermittent grazing with rest periods to permit recovery or frequent very light use. A lush growth of a palatable legume crop, such as white clover or alfalfa, often causes bloat in grazing animals, but this may be prevented by restricting intake. The use of drenches to prevent or break up foaming in the stomach act as temporary cures.

Alfalfa is the most important legume forage crop in the United States; it is used mainly for hay but is often grazed. White clover and the annual lespedezas are the most extensively grown legumes for grazing particularly in the southeastern United States. Red clover was an important crop prior to 1930 but is minor now. About a dozen other species of legumes are used for cultivated forage in the United States, and a large number are grown for range grazing. *See* ALFALFA; CLOVER; COVER CROPS; COWPEA; KUDZU; LESPEDEZA; LUPINE. Paul Tabor

Leitneriales

An order of flowering plants, division Magnoliophyta (Angiospermae) in the subclass Hamamelidae of the class Magnoliopsida (dicotyledons). The order consists of a single family, genus, and species (*Leitneria floridana*) of the southeastern United States. The plants are simple-leaved, dioecious shrubs with the flowers in catkins. The ovary is superior and pseudomonomerous, with a single ovule ripening into a small drupe. *See* HAMAMELIDAE; MAGNOLIOPHYTA; PLANT KINGDOM. Arthur Cronquist; T. M. Barkley

Lemming

The name applied to 11 species of rodents in the subfamily Microtinae, family Muridae. These animals (see **illustration**) have a northern circumpolar distribution. Lemmings have been extensively studied from the viewpoint of population dynamics, since their population increases cyclically, resulting in mass migration to the sea.

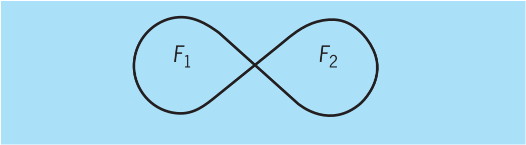
The lemming, a small burrowing rodent with short legs and stout claws adapted for digging. (*Photo by Robert R. Nack, Minnesota Cooperative Fish and Wildlife Research Unit*)

The Norway lemming (*Lemmus lemmus*) is found in the mountainous wastelands of northern Norway and Lapland. It is usually nocturnal and timid in its habits, except when a population explosion occurs with its resultant migration. Despite many reasons advanced to account for these migrations, the causative factors are not known. Cyclic variations in fertility may be a factor. Usually there are two litters of five offspring each year, but often four litters of two to eight offspring occur. The collared lemmings (*Dicrostonyx*) of Arctic Asia and America and the bog lemmings (*Synaptomys*) of North America have similar migrations. *Lemmus trimucronatus* is found in the Arctic regions of North America. It is a heavy-bodied, short-tailed rodent with short legs, large forefeet with a rudimentary thumb, and short hindfeet. The animal burrows extensively. This species has 16 teeth, with the dental formula I 1/1 C 0/0 Pm 0/0 M 3/3, and eats vegetation. The sexes are identical in appearance. The female may have from two to four litters a year with four to six offspring in each. *See* MUSKRAT; RODENTIA. Charles B. Curtin

Bibliography. R. M. Nowak, *Walker's Mammals of the World*, Johns Hopkins University Press, 1999; D. E. Wilson and S. Ruff, *The Smithsonian Book of North American Mammals*, 1999.

Lemniscate of Bernoulli

A curve shaped like the figure eight (see **illus.**), referred to by Jacques Bernoulli in 1694. Let F_1, F_2 be points of a plane π, with $F_1F_2 = 2a$, $a > 0$. The locus of a point P of π which moves so that $PF_1 \cdot PF_2 = b^2$, where b is a positive constant, is called an oval of Cassini. The lemniscate is obtained when $b = a$. Its equation in rectangular coordinates is $(x^2 + y^2)^2 = a^2(x^2 - y^2)$ and in polar coordinates $\rho^2 = a^2 \cos 2\theta$. It is the locus of the point of intersection of a variable tangent to a rectangular hyperbola with the line through the center perpendicular to the tangent. The area enclosed by the lemniscate $\rho^2 = a^2 \cos 2\theta$ is a^2. *See* ANALYTIC GEOMETRY. Leonard M. Blumenthal

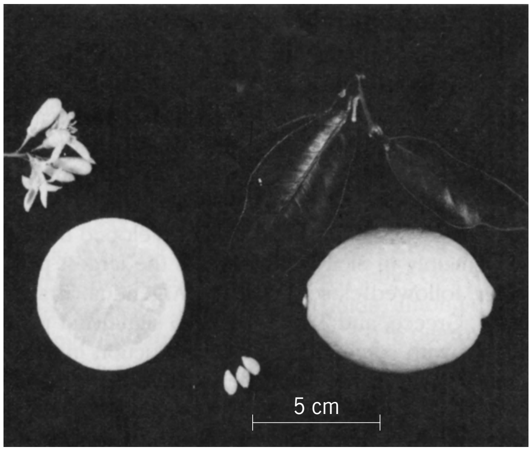


Curve known as a lemniscate.

Lemon

The fruit *Citrus limon*, commercially the most important of the acid citrus. It probably originated in the eastern Himalayan region of India or adjoining areas, and was introduced into Europe by the Arabs by the twelfth century.

Commercial lemon production developed first in Italy, mainly in Sicily. California is the largest producer, followed closely by Italy. Argentina, Brazil, Spain, Greece, and Turkey are also significant producers, with some commercial production in most Mediterranean countries and in Australia and South Africa. The United States produces about one-third of the total world production.



Flowers, foliage, fruit, and seeds of *Citrus limon*.

Culture is generally similar to that for other citrus fruits. Lemons are more sensitive to low temperatures than most citrus, and are susceptible to many problems in warm, humid climates. Therefore, commercial production is restricted to the warmer, drier portions of the citrus areas. In regions with mild winters and cool summers, marketable fruit is available throughout the year.

Lemons are grown on one of several rootstock types. None is entirely satisfactory, and the search for better ones has been ongoing. Sour orange (*C. aurantium*) has been widely used. Sweet orange (*C. sinensis*) is used, but is very susceptible to phytophthora root rot. In the United States, *C. macrophylla* is the most widely used rootstock. *Citrus macrophylla* trees are very productive but short-lived.

In the United States, the principal cultivars are largely classified into two groups, Eureka and Lisbon. Many selections have been made within each of these groups. In Italy, Feminello and selections of it are the most prominent. Berna is the most widely grown cultivar in Spain. Several other cultivars are grown to a limited extent. Most cultivars are vigorous, upright-spreading, and open in growth habit, attaining large size if not controlled by pruning. New shoot growth is purple-tinted, but mature leaves are a light green in contrast to the dark green of oranges. The large, purple-tinted flowers are produced throughout the year. The yellow fruits are medium-sized and elongate with a prominent nipple (see **illus.**).

The principal fungus diseases in the orchard are mal secco (caused by *Phoma tracheiphila*), which is a serious problem in the Mediterranean area, and brown rot (caused by *Phytophthora* sp.), which is a worldwide problem. Several virus diseases are spread through asexual propagation. Exocortis, a viroid, is also spread by pruning and is present in many cultivars. Insect pests are many and varied; several types of scale insects are serious problems, as are various mites. *See* PLANT PATHOLOGY.

The lemon is grown primarily for its acid flavor. For the fresh-fruit market, lemons are picked relatively immature and stored at 55–57°F (13–14°C) and 86–88% relative humidity to provide a constant supply of high-quality fruit. During storage, several fungi can cause serious losses. Blue and green molds (caused by *Penicillium* sp.) and sour rot (caused by *Geotrichum candidum*) are among the most troublesome. Others are *Alternaria* rot, brown rot, stem end rot, and *Trichoderma* rot. Tree-ripe fruits are processed because of their short storage life.

Lemon juice is very high in vitamin C. It has many culinary uses, especially in pies, cakes, ices, candies, jellies, and marmalades, and is also used widely in soft drinks. The principal by-product is lemon oil from the peel. *See* ASCORBIC ACID; FRUIT.

R. K. Soost

Length

A scalar physical quantity indicating a one-dimensional extension in space. Length is one of the three fundamental physical quantities important in mechanics, the other two being mass and time. It can be measured by comparison with an arbitrary standard; the specific one in common usage is the international meter. The most recent definition of the meter was adopted in 1983 at the meeting of the General Conference on Weights and Measures (Conférence Général des Poids et Mesures, or CPGM), when the meter was redefined in terms of time and the speed of light: “The meter is the length of the path traveled by light in a vacuum during a time interval of 1/299,792,458 of a second.” The effect was to take the speed of light in a vacuum to be exactly 299,792,458 m/s and then to define the meter in terms of that speed and the most accurately known quantity, the second. *See* LIGHT; TIME.

In theory, anyone with an accurate clock could then measure a meter. Such a time-of-flight method is impractical for most uses, and the meter is measured in terms of the wavelengths of light. The most accurate methods of determining lengths involve the interference of light waves. Distance can be measured by counting the static interference fringes in an interferometer to a precision of a fraction of the wavelength of the light. The wavelength of the light depends on the frequency of the light and on the index of refraction of the medium through which the light travels. The light sources used for metrology have well-known stable wavelengths with uncertainties as small as 1×10^{-10}. For measurements made in air, the index of refraction of the air also must be known to a high degree of precision. *See* FREQUENCY MEASUREMENT; INTERFEROMETRY; LASER; LASER SPECTROSCOPY; REFRACTION OF WAVES.

The meter, originally intended to be 1/10,000,000 of the distance from the Equator to the North Pole along a line of longitude through Paris, was defined by the first CPGM in 1889. It was defined as the distance between two fine lines engraved near the ends of a platinum-iridium bar when the bar was maintained at 0°C (32°F). The 11th CPGM (1960) replaced the prototype-bar standard with one based on the wavelength of light. Ultimately, measurements of the speed of light were limited by the precision of determining the meter. Finally, the CPGM accepted the best value of the speed of light as a fundamental constant and used that value to define the meter. *See* FUNDAMENTAL CONSTANTS.

Length bars and long gauge blocks are among the commonly used material standards that provide traceability to the definition of the meter. They are used to calibrate a wide range of instruments and other standards. The gauge blocks are calibrated by optical interferometry and may be used in combination to determine larger distances with precision. Longer bars and tapes are calibrated by comparison to these standards. Decimal multiples and submultiples of the meter are used to specify length. In the English system, the inch is defined to be exactly 0.0254 meter. *See* PHYSICAL MEASUREMENT.

Edwin R. Jones; Donald A. Jennings

Bibliography. Documents Concerning the New Definition of the Meter, *Metrologia*, vol. 19, 1984; E. R. Jones and R. L. Childers, *Contemporary College Physics 2001 Update*, 3d ed., McGraw-Hill, New York, 2001; B. N. Taylor, *The International System of Units (SI)*, NIST Spec. Publ., 330, 2001.

Lens (optics)

A curved piece of ground and polished or molded material, usually glass, used for the refraction of light. Its two surfaces have the same axis. Usually this is an axis of rotation symmetry for both surfaces; however, one or both of the surfaces can be toric, cylindrical, or a general surface with double symmetry. The intersection points of the symmetry axis with the two surfaces are called the front and back vertices and their separation is called the thickness of the lens.

Lens Types

There are three lens types, namely, compound, single, and cemented. These are described in the following sections.

Compound lenses. A compound lens is a combination of two or more lenses in which the second surface of one lens has the same radius as the first surface of the following lens and the two lenses are cemented together. Compound lenses are used instead of single lenses for color correction, or to introduce a surface which has no effect on the aperture rays but large effects on the principal rays, or vice versa. Sometimes the term compound lens is applied to any optical system consisting of more than one element, even when they are not in contact.

A group of lenses used together is a lens system. A symmetrical lens is a lens system consisting of two parts, each of which is the mirror image of the other. If one part is a mirror image of the other magnified m times, the system is called hemisymmetric. When $m = 1$, the system is often said to be holosymmetric.

Single lenses. The lens diameter is called the linear aperture, and the ratio of this aperture to the focal length is called the relative aperture. This latter quantity is more often specified by its reciprocal, called the *f*-number. Thus, if the focal length is 2 in. (50 mm) and the linear aperture 1 in. (25 mm), the relative aperture is 0.5 and the *f*-number is *f*/2. *See* FOCAL LENGTH.

In precalculation formulas, the lens thicknesses (but not the separations of the lenses) can frequently be neglected. This leads to the convenient fiction of a thin lens.

If ρ_1 and ρ_2 are the front and back curvatures of a lens of refractive index n and thickness d in air, its power is given by Eq. (1). The curvature of the surface is the reciprocal of its radius.

$$\phi = (n-1)(\rho_1 - \rho_2) + \frac{d(n-1)^2}{n}\rho_1\rho_2 \qquad (1)$$

The distances from the back vertex to the back nodal point and to the back focal point, respectively, are given by Eqs. (2). The last distance is often called the back focus, especially in photographic optics.

$$\begin{aligned} S'_N\phi &= -\frac{d}{n}(n-1)\rho_1 \\ S'_F\phi &= 1 + S'_N\phi \end{aligned} \qquad (2)$$

The bending of a lens is a change in the curvature of the two surfaces by the same amount. It does not change the power of a thin lens, which is $(n-1)$ $(\rho_1 - \rho_2)$. Bending is an important tool of the designer, for it permits the replacing of one lens by another without changing the data of gaussian optics.

When thick lenses are involved, gaussian optics remains constant only if both the powers of the thick lenses and the distance between the back nodal point of the first lens and the front nodal point of the second remain unchanged. Thus a bending of a thick lens should be accompanied by such an adjustment.

The optical center of a thick lens is the image of the nodal point produced inside the lens. All finite rays through the optical center emerge parallel to their respective directions at their entrance.

An optical center exists also in a hemisymmetric system. It is the point of symmetry which divides the separation of the two parts in the ratio $1/m$. If negative values of m are permitted, any single lens is a hemisymmetric system and the point dividing the thickness of the lens (the separation of the two vertices) in a ratio equal to the ratio of the two radii is the optical center of the lens.

A lens is said to be a collecting lens if $\phi > 0$ and a diverging lens if $\phi < 0$. When $\phi = 0$, the lens is afocal. Several types of collecting and diverging lenses are shown in **Fig. 1**.

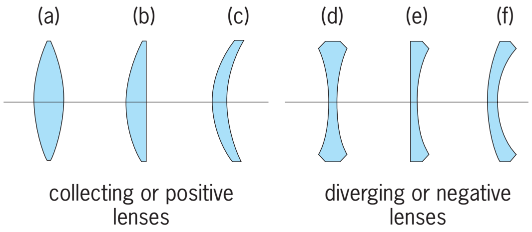


Fig. 1. Common lenses. (*a*) Biconvex. (*b*) Plano-convex. (*c*) Positive meniscus. (*d*) Biconcave. (*e*) Plano-concave. (*f*) Negative meniscus. (*After F. A. Jenkins and H. E. White, Fundamentals of Optics, 3d ed., McGraw-Hill, 1957*)

The surfaces of most lenses are either spherical or planar, but nonspherical surfaces are used on occasion to improve the corrections without changing the power of the lens. *See* OPTICAL SURFACES.

A concentric lens is a lens whose two surfaces have the same center. If the object to be imaged is also at the center, its axis point is sharply imaged upon itself, and since the sine condition is fulfilled, the image is free from asymmetry. Such a lens can be used as an additional system to correct meridional errors.

Another type of lens consists of an aplanatic surface followed by a concentric surface, or vice versa. Such a lens divides the focal length of the original lens to which it is attached by n^2, thus increasing the *f*-number by a factor of n^2 without destroying the axial correction of the preceding system. It does introduce curvature of field which makes a rebalancing of the whole system desirable. *See* ABERRATION (OPTICS).

Cemented lenses. Consider a compound lens made of two or more simple thin lenses cemented together. Let the power of the kth simple lens be Φ_k and its Abbe value ν_k. The difference between the powers of the combination for wavelengths corresponding to C and F is given by Eq. (3) where N may be considered

$$\Phi_F - \Phi_C = \frac{\Phi}{N} = \frac{\Sigma\,\Phi_k}{\nu_k} \qquad (3)$$

to be the effective ν-value of the combination. The ν-values of optical glasses vary between 25 and 70, with the ν-value of fluorite being slightly larger ($\nu = 95.1$). By using compound lenses, effective values of N can be obtained outside this range. Color correction is achieved as N becomes infinite, so that $\Phi_F - \Phi_C = 0$. A lens so corrected is called an achromat. In optical design, it is sometimes desirable to have negative values of N to balance the positive values of the rest of the system containing collecting lenses. Such a lens is said to be hyperchromatic. A cemented lens corrected for more than two colors is said to be apochromatic. A lens corrected for all colors of a sizable wavelength range is called a superachromatic lens. *See* CHROMATIC ABERRATION; OPTICAL MATERIALS.

Lens Systems

Optical systems may be divided into four classes: telescopes, oculars (eyepieces), photographic objectives, and enlarging lenses. *See* EYEPIECE; OPTICAL MICROSCOPE.

Telescope systems. A lens system consisting of two systems combined so that the back focal point of the first (the objective) coincides with the front focal point of the second (the ocular) is called a telescope. Parallel entering rays leave the system as parallel rays. The magnification is equal to the ratio of the focal length of the first system to that of the second.

If the second lens has a positive power, the telescope is called a terrestrial or keplerian telescope and the separation of the two parts is equal to the sum of the focal lengths.

If the second lens is negative, the system is called a galilean telescope and the separation of the two parts is the difference of the absolute focal lengths. The galilean telescope has the advantage of shortness (a shorter system enables a larger field to be corrected); the keplerian telescope has a real intermediate image which can be used for introducing a reticle or a scale into the intermediate plane.

Both objective and ocular are in general corrected for certain specific aberrations, while the other abberations are balanced between the two systems. *See* TELESCOPE.

Photographic objectives. A photographic objective images a distant object onto a photographic plate or film. *See* PHOTOGRAPHY.

The amount of light reaching the light-sensitive layer depends on the aperture of the optical system, which is equivalent to the ratio of the lens diameter to the focal length. Its reciprocal is called the *f*-number. The smaller the *f*-number, the more light strikes the film. In a well-corrected lens (corrected for aperture and asymmetry errors), the *f*-number cannot be smaller than 0.5.

The larger the aperture (the smaller the *f*-number), the less adequate may be the scene luminance required to expose the film. Therefore, if pictures of objects in dim light are desired, the *f*-number must be small. On the other hand, for a lens of given focal length, the depth of field is inversely proportional to the aperture.

Since the exposure time is the same for the center as for the edge of the field, it is desirable for the same amount of light to get to the edge as gets to the center, that is, the photographic lens should have little vignetting.

The camera lens can be considered as an eye looking at an object (or its image), with the diaphragm corresponding to the eye pupil. The gaussian image of the diaphragm in the object (image) space is called the entrance (exit) pupil. The angle under which the object (image) is seen from the entrance (exit) pupil is called the object (image) field angle. For most photographic lenses, the entrance and exit pupils are close to the respective nodal points; for such lenses, the object and the image field angles are equal.

In general, photographic objectives with large fields have small apertures: those with large

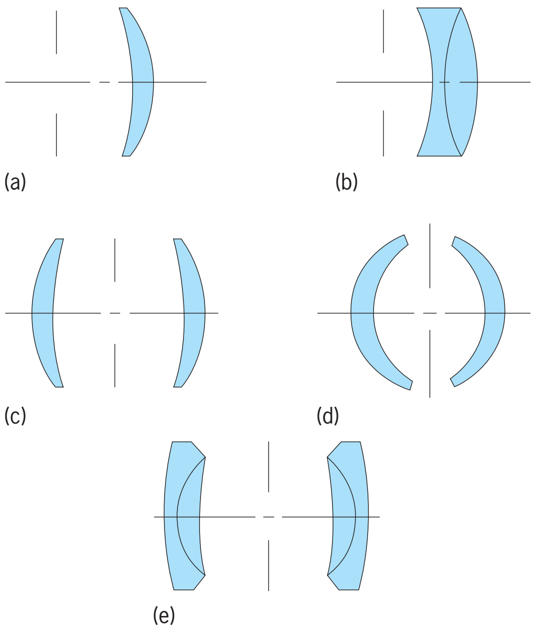


Fig. 2. Older cameralenses. (*a*) Meniscus. (*b*) Simple achromat. (*c*) Periskop. (*d*) Hypergon wide-angle. (*e*) Symmetrical achromat.

apertures have small fields. The construction of the two types of systems is quite different. One can say in general that the larger the aperture, the more complex the lens system must be.

There exist cameras (so-called pinhole cameras) that do not contain any lenses. The image is then produced by optical projection. The aperture in this case should be limited to *f*/22.

Other types of lenses. A single meniscus lens, with its concave side toward the object and with its stop in front at its optical center, gives good definition at *f*/16 over a total field of 50° (**Fig 2***a*). The lens can be a cemented doublet for correcting chromatic errors (Fig. 2*b*). For practical reasons, a reversed meniscus with the stop toward the film is often used.

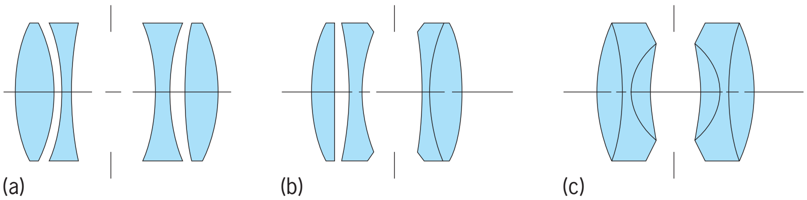


Fig. 3. Types of anastigmats. (*a*) Celor. (*b*) Tessar. (*c*) Dagor.

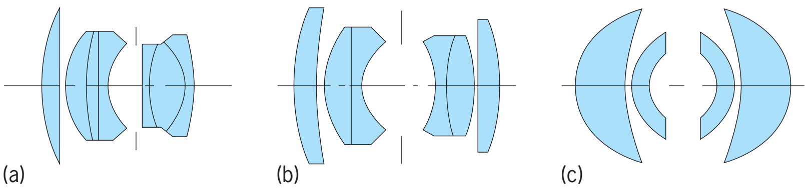


Fig. 4. Modern camera lenses. (*a*) Sonnar. (*b*) Biotar. (*c*) Topogon.

Combining two meniscus lenses to form a symmetrical lens with central stop makes it possible to correct astigmatic and distortion errors for small apertures as well as large field angles (Fig. 2*c*).

The basic type of wide-angle objective is the Hypergon, consisting of two meniscus lenses concentric with regard to the stop (Fig. 2*d*). This type of system can be corrected for astigmatism and field curvature over a total field angle of 180° but it can only be used for a small aperture (*f*/12), since it cannot be corrected for aperture errors. The aperture can be increased to *f*/4 at the expense of field angle by thickening and achromatizing the meniscus lenses and adding symmetrical elements in the center or at the outside of the basic elements.

Two positive achromatic menisci symmetrically arranged around the stop led to the aplanatic type of lens (Fig. 2*e*). This type was spherically and chromatically corrected. Since the field could not be corrected, a compromise was achieved by balancing out sagittal and meridional field curvature so that one image surface lies in front and the other in back of the film.

Anastigmatic lenses. The discovery of the Petzval condition for field correction led to the construction of anastigmatic lenses, for which astigmatism and curvature of field are corrected. Such lenses must contain negative components.

The Celor (Gauss) type consists of two air-spaced achromatic doublets, one on each side of the stop (**Fig. 3***a*). The Cooke triplet combines a negative lens at the aperture stop with two positive lenses, one in front and the other in back. It is called a Tessar (Fig. 3*b*) if the last positive lens is a cemented doublet, or a Heliar if both positive lenses are cemented. The Dagor type consists of two lens systems that are nearly symmetrical with respect to the stop, each system containing three or more lenses (Fig. 3*c*).

Modern lenses. To increase the aperture, the field, or both, it is frequently advantageous to replace one lens by two separated lenses, since the same power is then achieved with larger radii and this means that the single lenses are used with smaller relative apertures. The replacing of a single lens by a cemented lens changes the color balance, and thus the designer may achieve more favorable conditions. Moreover, the introduction of new types of glass (first the glasses containing barium, later the glasses containing rare earths) led to lens elements which for the same power have weaker surfaces and are of great help to the lens designer, since the errors are reduced.

Of modern designs the most successful are the Sonnar, a modified triplet, one form of which is shown in **Fig. 4***a*; the Biotar (Fig. 4*b*), a modified Gauss objective with a large aperture and a field of about 24°; and the Topogon (Fig. 4*c*), a periscopic lens with supplementary thick menisci to permit the correction of aperture aberrations for a moderate aperture and a large field. One or two plane-parallel plates are sometimes added to correct distortion.

Special objectives. It is frequently desirable to change the focal length of an objective without changing the focus. This can be done by combining a fixed near component behind the stop with an exchangeable set of components in front of the stop. The designer has to be sure that the errors of the two parts are balanced out regardless of which front component is in use. For modern ways to change the magnification *see* ZOOM LENS

The telephoto objective is a specially constructed objective with the rear nodal point in front of the lens, to combine a long focal length with a short back focus. *See* TELEPHOTO LENS.

The Petzval objective is one of the oldest designs (1840) but one of the most ingenious. It consists in general of four lenses ordered in two pairs widely separated from each other. The first pair is cemented and the second usually has a small air space. For a relatively large aperture, it is excellently corrected for aperture and asymmetry errors, as well as for chromatic errors and distortion. It is frequently used as a portrait lens and as a projection lens because of its sharp central definition. Astigmatism can be balanced but not corrected.

Enlarger lenses and magnifiers. The basic type of enlarger lens is a holosymmetric system consisting of two systems of which one is symmetrical with the first system except that all the data are multiplied by the enlarging factor *m*. When the object is in the focus of the first system, the combination is free from all lateral errors even before correction. A magnifier in optics is a lens that enables an object to be viewed so that it appears larger than its natural size.

The magnifying power is usually given as equal to one-quarter of the power of the lens expressed in diopters. *See* DIOPTER; MAGNIFICATION.

Magnifying lenses of low power are called reading glasses. A simple planoconvex lens in which the principal rays are corrected for astigmatism for a position of the eye at a distance of 10 in. (25 cm) is well suited for this purpose, although low-power magnifiers are often made commercially with biconvex lenses. A system called a verant consists of two lenses corrected for color, astigmatism, and distortion. It is designed for stereoscopic vision at low magnification. *See* STEREOSCOPY.

For higher magnifications, many forms of magnifiers exist. One of the basic designs has the form of a full sphere with a diaphragm at the center, as shown in **Fig. 5***a*. The sphere may be solid or it may be filled with a refracting liquid. When it is solid, the diaphragm may be formed by a deep groove around the equator. Combinations of thin planoconvex lenses as shown in Fig. 5*b* and *c* are much used for moderate powers. Better correction can be attained in the aplanatic magnifier of C. A. Steinheil, in which a biconvex crown lens is cemented between a pair of flint lenses (Fig. 5*d*).

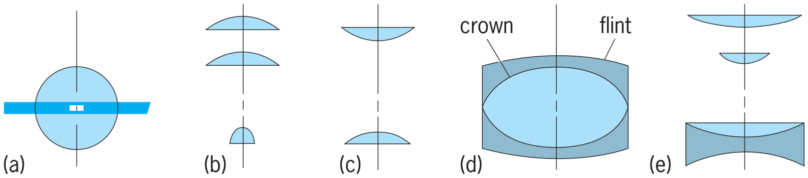


Fig. 5. Typical magnifiers. (*a*) Sphere with equatorial diaphragm. (*b*,*c*) Planoconvex lens combinations. (*d*) Steinheil triple aplanat. (*e*) Chevalier type.

A design by C. Chevalier (Fig. 5*e*) aims for a large object distance. It consists of an achromatic negative lens combined with a distant collecting front lens. A magnifying power of up to 10× with an object distance up to 3 in. (75 mm) can be attained.

Max Herzberger

Bibliography. E. Hecht and A. Zajac, *Optics*, 3d ed., 1997; M. Herzberger, *Modern Geometrical Optics*, 1958, reprint 1980; F. A. Jenkins and M. E. White, *Fundamentals of Optics,* 4th ed., 1976; R. Kingslake, R. R. Shannon, and J. C. Wyant (eds.), *Applied Optics and Optical Engineering*, 15 vols., 1965–2000; D. Malacara and Z. Malacara, *Handbook of Lens Design*, 1994; S. F. Ray, *Applied Photographic Optics*, 2d ed., Focal Press, 1996; W. J. Smith, *Modern Lens Design: A Resource Manual*, 1992.

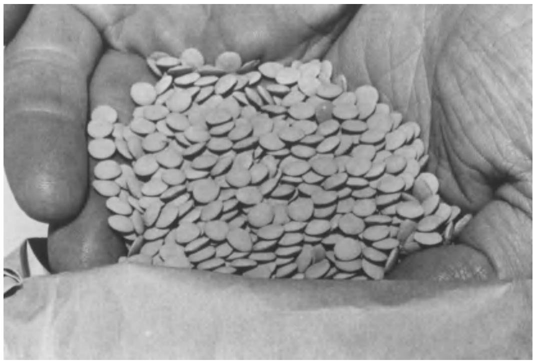
Lentil

A semi-viny annual legume with slender tufted and branched stems 18–22 in. (46–56 cm) long. The lentil plant (*Lens esculenta*) was one of the first plants brought under cultivation. They have been found in the Bronze Age ruins of the ancient lake dwellings of St. Peter's Island, Lake of Bienne, Switzerland. Lentils have been discovered in Poland dating back to the Iron Age. In the Bible the "red pottage" for which Esau gave up his birthright to his brother, Jacob, was probably lentil soup. Large-seeded lentils originated in the Mediterranean region; medium-sized lentils originated in the inner mountains of Asia Minor; and Afghanistan was the original home of the smallest-seeded lentils. *See* LEGUME.

Production. The world's lentil production is centered in Asia, with nearly two-thirds of the production from India, Pakistan, Turkey, and Syria. Whitman and Spokane counties in Washington, and Latah, Benewah, and Nez Perce counties in Idaho grow about 95% of the lentils produced in the United States.

Description. Lentil leaves are pinnately compound and generally resemble vetch. The plant has tendrils similar to those of pea plants.

The seeds grow in short broad pods, each pod producing two or three thin lens-shaped seeds (see **illus.**). Seed color varies from yellow to brown and may be mottled, although mottled seeds are not desirable for marketing. The lentil flowers, ranging in color from white to light purple, are small and delicate and occur at different locations on the stems.



Lens-shaped lentil seeds.

Lentil seed is used primarily for soups but also in salads and casseroles. Lentils are more digestible than meat and are used as a meat substitute in many countries.

Culture. Lentils require a cool growing season; they are injured by severe heat. Therefore, they are planted in April, when soil moisture is adequate and temperatures are cool. A fine firm seedbed is required; the land is usually plowed in the fall and firmed by cultivation in the spring before seeding. Lentils are usually planted in rotation with winter wheat. Seeds are planted in 7–12 in. (21–30 cm) rows at depths of $^{1}/_{2}$ to 1 in. (13 to 25 mm), on an average of 60–75 lb/acre (67–84 kg/hectare). Applications of sulfur, molybdenum, and phosphorus are used to increase yields.

Wild oats often infest lentil fields, but chemicals are available to aid weed control. Cowpea and black-bean aphids are the two most important insect pests on lentils. Predators such as ladybird beetles and syrphid-fly larvae usually keep these insects under control, but when their populations are too low to control the aphids, chemical insecticides are used.

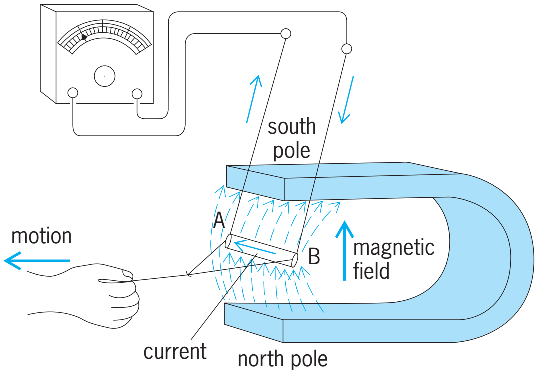
Harvesting. Lentils are mowed or swathed when the vines are green and the pods have a golden color. About 10 days later, lentils are ready to combine harvest using the same combines that are used for wheat, oats, and barley. A pick-up attachment picks up the material from the windrows. The combines must be operated at a maximum speed of $1^{1}/_{2}$ mi/h (0.7 m/s) to prevent loss of, and damage to, the lentil seed. Average yields are about 900 lb/acre (1000 kg/ha). *See* AGRICULTURAL MACHINERY; AGRICULTURAL SOIL AND CROP PRACTICES.

Kenneth J. Morrison

Bibliography. O. N. Allen and E. K. Allen, *The Leguminosae: A Source Book of Characteristics, Uses, and Nodulation*, 1981; R. J. Summerfield and A. H. Bunting (eds.), *Advances in Legume Science*, vol. 1, 1980.

Lenz's law

A law of electromagnetism which states that, whenever there is an induced electromotive force (emf) in a conductor, it is always in such a direction that the current it would produce would oppose the change which causes the induced emf. If the change is the motion of a conductor through a magnetic field, as in the **illustration**, the induced current must be in such a direction as to produce a force opposing the motion. If the change causing the emf is a change of flux threading a coil, the induced current must produce a flux in such a direction as to oppose the change. That is, if the change is an increase of flux, the flux due to the induced current must be opposite in direction to the increasing flux. If the change is a decrease in flux, the induced current must produce flux in the same direction as the decreasing flux.



Induced emf in a moving conductor. Direction of current induced in wire AB is indicated by the arrows. (*After M. W. White, K. V. Manning and R. L. Weber, Practical Physics, 2d ed., McGraw-Hill, 1955*)

Lenz's law is a form of the law of conservation of energy, since it states that a change cannot propagate itself. *See* CONSERVATION OF ENERGY; ELECTROMAGNETIC INDUCTION; ELECTROMOTIVE FORCE (EMF).

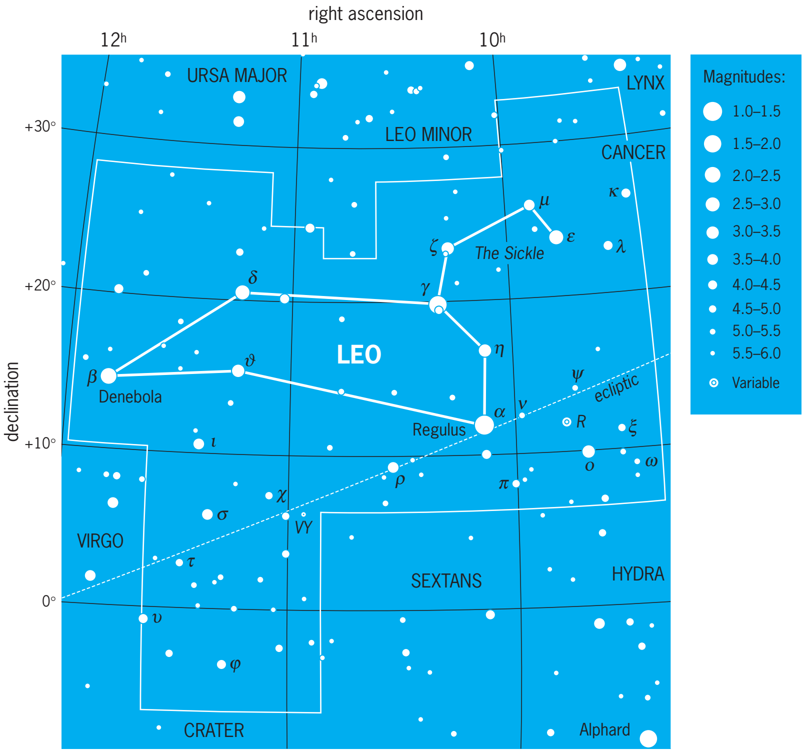
Kenneth V. Manning

Leo

The Lion, a northern zodiacal constellation (see **illustration**). Its head and front legs are marked by stars tracing out a backward question mark (without a bottom dot), also known as the Sickle. The bright star Regulus is at the bottom of the Sickle. The end of Leo's tail is marked by the bright star Denebola; its name is Arabic for lion's tail. *See* ZODIAC.

Leo is in the evening sky during the spring and early summer. The Leonid meteor shower, which appears to emanate from Leo each November 17, is especially outstanding every 33 years, most recently in 1999–2001, so it will next be prominent starting around 2032. *See* METEOR.

The modern boundaries of the 88 constellations,



Modern boundaries of the constellation Leo, the Lion. The celestial equator is 0° of declination, which corresponds to celestial latitude. Right ascension corresponds to celestial longitude, with each hour of right ascension representing 15° of arc. Apparent brightness of stars is shown with dot sizes to illustrate the magnitude scale, where the brightest stars in the sky are 0th magnitude or brighter and the faintest stars that can be seen with the unaided eye at a dark site are 6th magnitude. (*Wil Tirion*)

including this one, were defined by the International Astronomical Union in 1928. *See* CONSTELLATION.

Jay M. Pasachoff

Lepidodendrales

An extinct order of the class Lycopsida which, together with Isoetales, forms the monophyletic rhizomorphaleans, the most derived and diverse group of clubmosses. The lepidodendraleans are best known as the scale trees that dominate most reconstructions of Upper Carboniferous swamps, where they were the main constituent of many coal-forming peats.

Lepidodendraleans are represented most frequently in the fossil record as characteristic “tire-track” bark fragments, which reflect the regular geometric arrangements of leaves, although taxonomic revisions have focused on the more complete information obtained from fossils preserved three-dimensionally in petrified peats. Reconstruction of the plants from their component organs has revealed that the classic trees (**Fig. 1**) possessed a wide range of growth architectures and shared lowland habitats with other smaller-bodied genera, reflecting a major radiation that occurred in the Late Devonian and perhaps Early Carboniferous.

Rhizomorphic lycopsids are distinguished by their centralized rootstock (rhizomorph), which permitted bipolar growth; both aerial and subterranean axes could branch, unlike the unipolar rhizomes of other, more primitive lycopsids. This trait and the newly acquired ability to produce wood allowed many rhizomorphic lycopsids to form large upright trees, with trunks up to 115 ft (35 m) tall and 3 ft (1 m) in diameter. The tree-lycopsids relied primarily on the external cylinder of bark for support rather than the relatively narrow central cylinder of wood, which was adapted primarily for water transport. Photosynthesis occurred largely in the leaf bases and bark, as well as in the narrow, grasslike leaf laminae.

The lepidodendraleans consisted of four main growth modules: rhizomorph, stem, lateral branches, and crown branches. Only one of the two branch types was well expressed in any one genus. The growth of the tree-lycopsids was probably very rapid. Trees with disposable lateral branches reached reproductive maturity earlier than those that relied only on crown branches. However, unlike most seed plants, branches functioned primarily for reproduction and propagule dispersal rather than light capture; even when mature, they formed open canopies that cast little shade.

The evolution of the group demonstrably entailed progressively increased reproductive sophistication. The more primitive genera possessed bisporangiate cones resembling those of the Selaginellales (**Fig. 2**). Typically, megasporangia containing “female” megaspores developed at the base of the cone where greater resources were available, and

Fig. 1. **Reconstructions of tree-sized Carboniferous rhizomorphic tree-lycopsids, illustrating architectural variation. (*After R. M. Bateman et al., Experimental cladistic analysis of anatomically preserved arborescent lycopsids from the Carboniferous of Euramerica, Ann. Mo. Bot. Gard., 79:500–559, 1992*)**

microsporangia containing "male" microspores developed closer to the cone apex. The genus *Sigillaria* is characterized by the segregation of the two genders of sporangium in different cones, allowing modifications of the female cone without necessarily modifying the male cone. All subsequent genera underwent selective abortion to leave only one functional megaspore in each megasporangium, now having thinner walls and lacking external ornamentation since it could no longer serve any function. The sporophylls then enclosed the megasporangium to form dispersal units that were the largest of any pteridophyta. Termed aquacarps, they typify *Lepidophlois* and *Miadesemia*. *See* SELAGINELLALES.

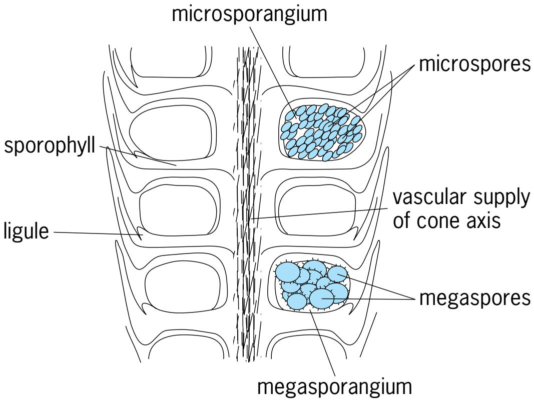


Fig. 2. **Primitive bisexual cone of the *Flemingites* type, borne by *Paralycopodites*, with apically concentrated microsporangia and basally concentrated megasporangia containing several megaspores. (*After C. A. Arnold, An Introduction to Paleobotany, McGraw-Hill, 1947*)**

These reproductive changes were accompanied by vegetative modifications. Notably, stelar morphology and leaf base anatomy (the basis of a separate organ taxonomy for bark) became more complex, particularly in the larger genera, and ligules became recessed in pits. In contrast, no such progressive increase in evolutionary sophistication occurred in the gross vegetative morphology of the plants. Frequent, profound, and broadly heterochronic changes in body size favored transitions from trees to shrubs and pseudoherbs (plants that were woody and determinate but small bodied and recumbent).

The lepidodendraleans partitioned and occupied most potential ecological niches, at least in the extensive European-American swamps of the Late Carboniferous. The warming and drying of the global climate toward the end of this period greatly reduced the diversity of the rhizomorphic lycopsids, beginning with the most derived and specialized genera such as *Lepidophloios* and eventually leaving only the increasingly reduced and ecologically specialized isoetaleans to survive to present day. *See* CARBONIFEROUS; ISOETALES; LYCOPHYTA; LYCOPSIDA.

Richard M. Bateman; William A. DiMichele

Bibliography. R. M. Bateman, Evolutionary-developmental change in the growth architecture of fossil mizomorphic lycopsids, *Biol. Rev.,* 69:527–597; T. L. Phillips and W. A. DiMichele, Comparative ecology and life-history biology of arborescent lycopsids, *Ann. Mo. Bot. Gard.*, 79:560–588, 1992; W. N. Stewart and G. W. Rothwell, *Paleobotany and the Evolution of Plants*, 2d ed., 1993.

Lepidolite

A mineral of variable composition which is also known as lithium mica and lithionite, $K_2(Li,Al)_{5-6}(Si_{6-7},Al_{2-1})O_{20-21}(F,OH)_{3-4}$. Rubidium (Rb) and cesium (Cs) may replace potassium (K); small amounts of Mn, Mg, Fe(II), and Fe(III) normally are present; and the OH/F ratio varies considerably. Polithionite is a silicon- and lithium-rich, and thus aluminum-poor, variety of lepidolite.

Lepidolite is uncommon, occurring almost exclusively in structurally complex granitic pegmatites, commonly in replacement units. Common associates are quartz, cleavelandite, alkali beryl, and alkali tourmaline. Lepidolite is a commercial source of lithium, commonly used directly in lithium glasses and other ceramic products. Important deposits occur in the Karibib district of South-West Africa and at Bikita, Rhodesia.

The structural modifications show some correlation with lithium content: the six-layer monoclinic form contains 4.0–5.1% Li_2O; the one-layer monoclinic, 5.1–7.26% Li_2O. A three-layer hexagonal form is also found. There is a compositional gradation to muscovite, intermediate types being called lithian

Group of lepidolite crystals found in Zinnwald, Czechoslovakia. (*Specimen from Department of Geology, Bryn Mawr College*)

muscovite, containing 3–4% Li_2O, and having a modified two-layer monoclinic muscovite structure.

Lepidolite usually forms small scales or fine-grained aggregates (see **illus.**). Its colors, pink, lilac, and gray, are a function of the Mn/Fe ratio. It is fusible at 2, yielding the crimson (lithium) flame. It has a perfect basal cleavage. Hardness is 2.5–4.0 on Mohs scale; specific gravity is 2.8–3.0. *See* MICA; SILICATE MINERALS.

E. William Heinrich

Lepidoptera

The order of scaly-winged insects, including the butterflies, skippers, and moths. One of the largest orders in the class Insecta, the Lepidoptera include over 150,000 known species (of which about 10,000 occur in North America) divided among more than 100 families. The adults have a covering of hairs and flattened setae (scales) on the wings, legs, and body, and are often beautifully colored. With minor exceptions, the adults are also characterized by two pairs of membranous wings, and sucking mouthparts that feature a prominent coiled proboscis. This feeding apparatus is formed from a pair of specially elongated and grooved lobes, the galeae (a mouthpart), that are closely joined along their length to make a flexible tube. Adults having a proboscis (the vast majority) can take only liquid food, such as nectar and juices of fruits. Butterflies and skippers usually fly in the daytime, while most moths are nocturnal. The caterpillars are almost always herbivorous and chew their food like a grasshopper or beetle. *See* INSECTA.

Morphology. The head of most adult Lepidoptera species is dominated by the bulbous compound eyes (**Fig. 1***a*), as seen in other insects. The most unusual feature of most moths and butterflies, however, is the form of the mouthparts (Fig. 1*a–c*). The proboscis is thought to be extended by blood pressure created by retraction of special muscles at its base. Other muscles, arranged diagonally within each half of the proboscis, appear to be necessary for the whole proboscis to coil. Liquid is drawn up into the proboscis by the action of a muscular sucking pump inside the insect's head (**Fig. 2***a*). Simple eyes (ocelli) located near the top of the head (Fig. 1*b*) are absent in many groups, such as the Hesperioidea and Papilionoidea. The antennae are variable in form (**Fig. 3**).

Thorax. The prothorax, the first thoracic segment, is well developed in some lower groups, such as the Hepialoidea, but it is considerably smaller than the second and third pterothoracic, or wing-bearing, segments, and is largely membranous in the majority of Lepidoptera. The most prominent feature of the dorsum of the prothorax, in most groups, is a pair of protuberant lobes, the patagia.

The paired basal segments (coxae) of the prothoracic legs are cylindrical and functional, while the two pairs of coxae of the pterothorax (mesothorax and metathorax) are fused with the thoracic capsule and are thus immobile. In each segment of the thorax, the midventral suture, the discrimen, is represented internally by a fine membranous lamella. The form of this lamella and its relationship to the furca (a forked internal projection arising from the base of the thorax), especially in the mesothorax, have proved significant for the classification of some Lepidoptera, including butterflies. The mesothorax is the largest of the three segments and may completely overlap the metathorax.

There is considerable variation throughout the order in the details of the many different sclerites (the hardened plates that make up the body wall) and their connecting sutures. In the butterflies, interesting variation has also been found in the structure of the gut, but this has not been studied extensively in the rest of the order. Many questions remain unanswered concerning the morphology, function, and evolution of this fundamental body system.

Scales. The wing scales are very variable in form (Fig. 1*d*). Generally, they are flat, thin, cuticular sacks with striated upper surfaces. They have a basal stalk, the pedicel, that fits into a socket in the wing membrane. The males of many species have special scales called androconia. These have feathered tips (Fig. 1*e*) or various other forms, and are often gathered together as special androconial organs. Their usual function is to disseminate scents, notably those used by the males during courtship. The vast spectrum of colors seen in the Lepidoptera can be grouped into two sorts, pigmentary and structural. The former result from various chemical pigments, such as melanin (black/brown/orange) or papiliochrome (yellow), deposited in the scales. Structural colors are produced either by fine diffracting ridges on the surface of the scales or, more usually, by layers within the scales that are so close together they interfere with light. These diffraction or interference colors are most often metallic or iridescent greens and blues, but the remarkable cryptic green of some hairstreak butterflies (for example, *Callophrys rubi*) is structural, yet totally noniridescent. In some Lepidoptera, blue bile pigments deposited between the wing membranes also contribute to the overall color pattern, as in species of *Graphium* (Papilionidae).

Wings. In most moths, the fore- and hindwings on each side are coupled together. In females this is achieved by a group of stiff bristles or setae, or

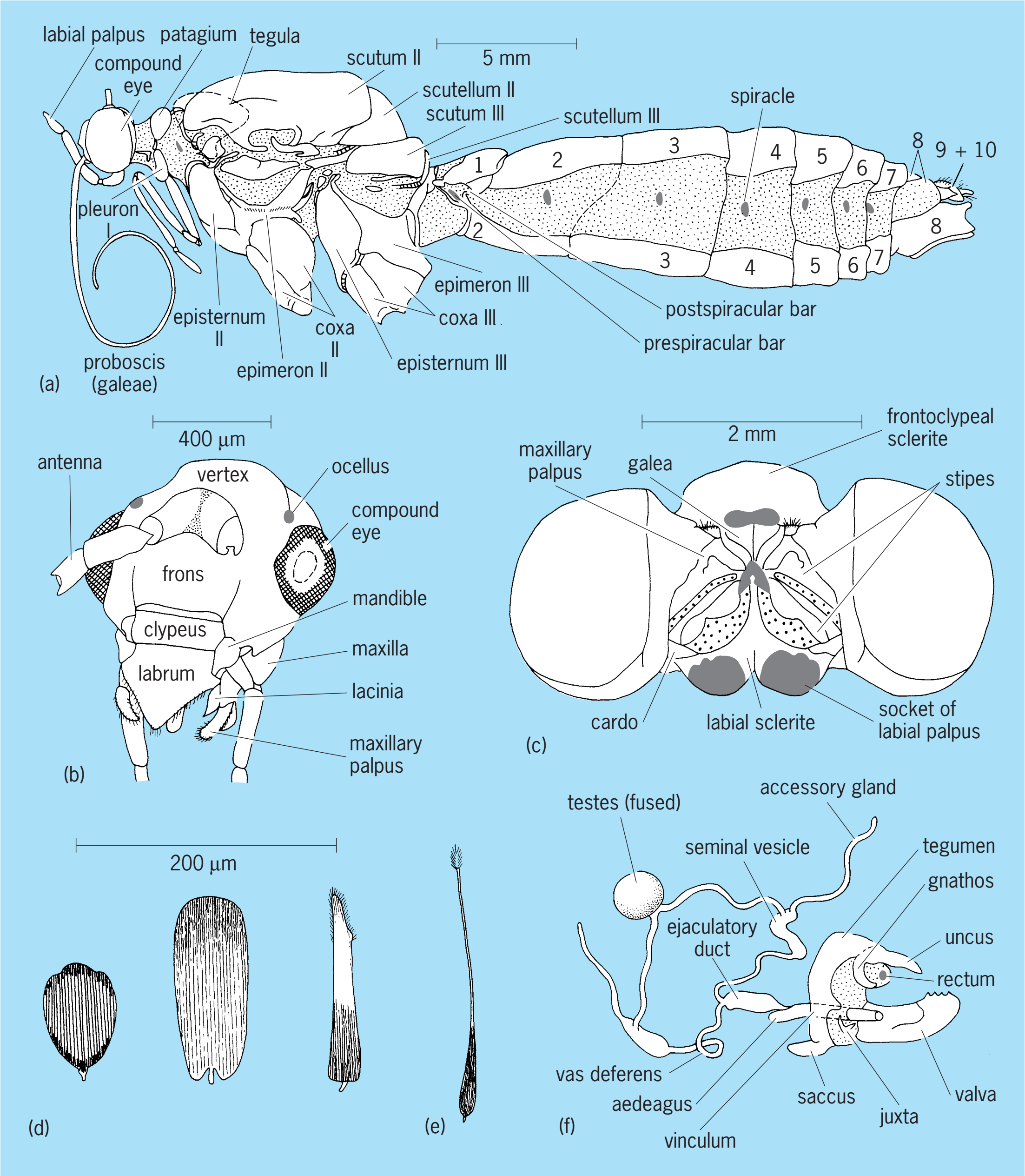


Fig. 1. Adult Lepidoptera anatomy. (*a*) Lateral view of *Danaus plexippus* (Nymphalidae) with hairs, scales, wings, and mid- and hindlegs removed. (*b*) Front view of head of *Epimartyria* (Micropterygidae). (c) Ventral view of head of *D. plexippus*. (*d*) Unspecialized scales. (e) Androconia (specialized male scales). (*f*) Male genital system.

in males by a large spine formed from several fused setae that projects forward from the base of the hindwing and is held by a clasp on the forewing. The spine or group of setae is known as the frenulum and the clasp as the retinaculum (**Fig.** 4*a*). In homoneurous moths (the shape and venation of the fore- and hindwings are similar; see below), there is a lobe, the jugum, at the base of the forewing, which engages with the hindwing or with the frenulum when it is present (Fig. 4*b*). In butterflies, skippers, and some moths, the humeral angle of the hindwing is expanded and strengthened by one or more humeral veins (Fig. 4*c*). In these groups, the frenulum is usually lost, the wings being united functionally by the overlapping lobe.

Genitalia. The external genitalia, together with the anus, usually occupy the last three segments of the abdomen. The form of the external genitalia, especially of the male (Fig. 1*f*), has been used widely in the separation and classification of species. Thus, the valva, uncus, gnathos, and saccus often differ widely between groups and sometimes even closely related species. During copulation, the vulva (ostium bursae) of the female receives the phallus (aedeagus) of the male, through which the sperm is passed in a packet, the spermatophore. The female stores the spermatophore in a special chamber, the corpus bursae, and can then fertilize her eggs as they are laid, sometimes many days later. This is achieved in most Lepidoptera by the sperm passing along a seminal duct that connects the corpus bursae with the egg duct (oviduct), through which the eggs must pass before they can be laid. In some groups the oviduct terminates in a distinct egg-laying apparatus, the ovipositor.

Developmental Stages

Development starts with the fertilized egg nucleus, which divides and grows, nourished by nutrients stored inside the egg, to give rise to the first larval stage (first instar). This has to bite its way out of the egg shell to commence its independent development. The larvae of Lepidoptera, commonly called caterpillars, are mandibulate and cylindrical, with short articulated thoracic legs and a variable number of fleshy abdominal prolegs. They have one pair of thoracic spiracles and eight pairs of abdominal spiracles (external breathing orifices). After the larva grows and sheds its skin several times, the skin of the fully grown, last larval instar is eventually shed to reveal the immobile pupa. This is the stage during which metamorphosis takes place, the transformation between the wormlike larval stage and winged adult. Inside the pupa, most of the old larval structures are broken down and used as materials and energy supplies to build the adult insect. The pupae of Lepidoptera are variable in form and are often enclosed in a silken cocoon. *See* INSECT PHYSIOLOGY; METAMORPHOSIS.

Egg. There are two general types of eggs, flattened and upright. In both types the surface of the egg shell (the chorion) is usually sculptured, with the sculpturing on the upright type generally more complex (**Fig. 5**). There is a microscopic entrance in the chorion, the micropyle, through which sperm can enter. The eggs are usually laid on the food plant, either singly or in clusters, attached by means of a glue or cement. However, the eggs are sometimes merely scattered in the vicinity of the food plant.

Larva. The caterpillars are variable in color, size, and shape; the presence of warts, hairs, and setae; the arrangement of hooks (or crochets) on the prolegs; and many other superficial details. They are rather constant, however, in basic morphology (**Fig. 6**). There are usually six pairs of ocelli, and compound eyes are absent. At most, eight pairs of prolegs are present, but almost every degree of reduction is found; most species have five pairs of prolegs. *See* CATERPILLAR.

Silk is produced by the labial glands. They are long, coiled, simple tubes which unite anteriorly in a cylindrical spinneret, opening at the front of the labium. *See* SILK.

Pupa. There are three main types, although the vast majority are obtect, with the appendages (legs, antennae, wings) completely sheathed with the rest of the body and immobile, and only the terminal abdominal segments movable. In the Hepialoidea, Cossoidea, and other lower moths, more of the abdominal segments are movable (**Fig. 7**). Although sheathed, the pupal appendages of the Hepialoidea and Cossoidea are partly free, and their pupae are referred to as incomplete. The

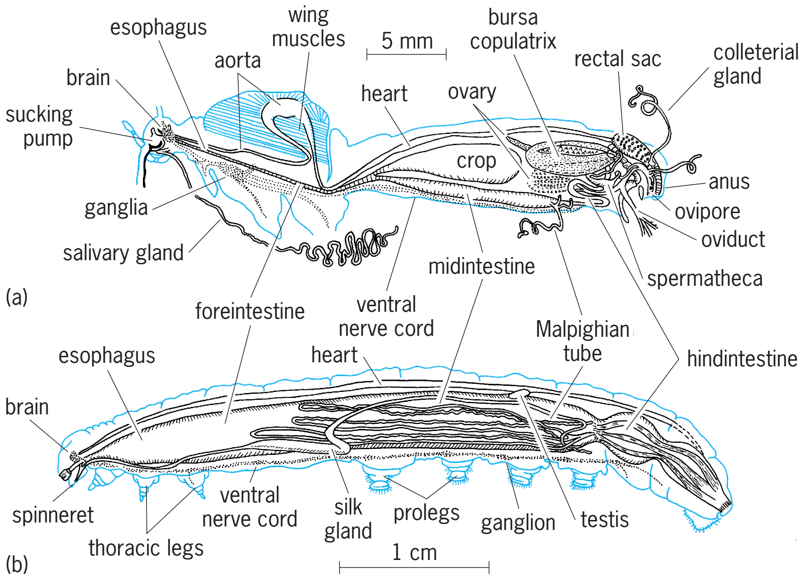


Fig. 2. Internal anatomy of *Danaus plexippus* (Nymphalidae) with tracheal (breathing tube) system and most of musculature omitted. (*a*) Adult female. (*b*) Larval male.

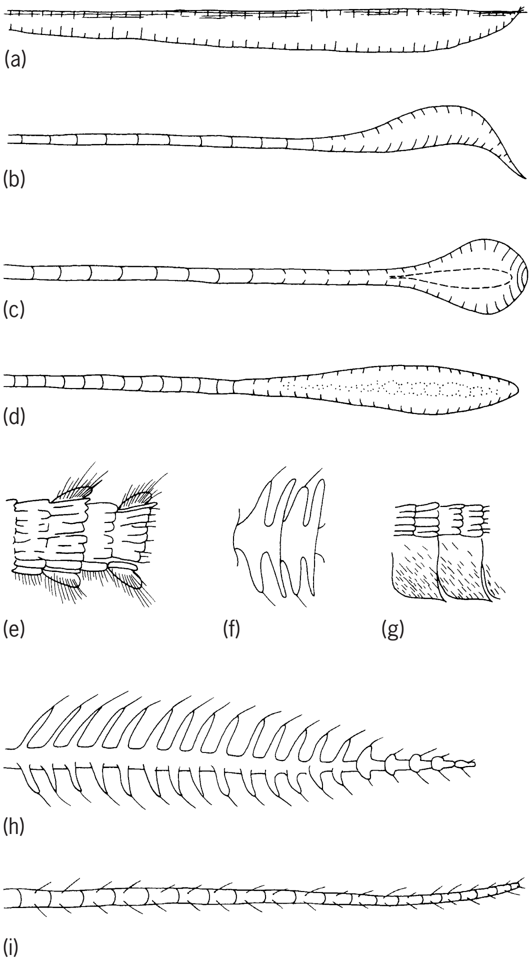


Fig. 3. Types of Lepidoptera antennae. (*a*) Fusiform. (*b*) Clubbed and hooked. (*c*, *d*) Clubbed. (*e*) Serrate and fasciculate, dorsal view. (*f*) Doubly bipectinate (quadripectinate). (*g*) Laminate, lateral view. (*h*) Bipectinate. (*i*) Simple and ciliate.

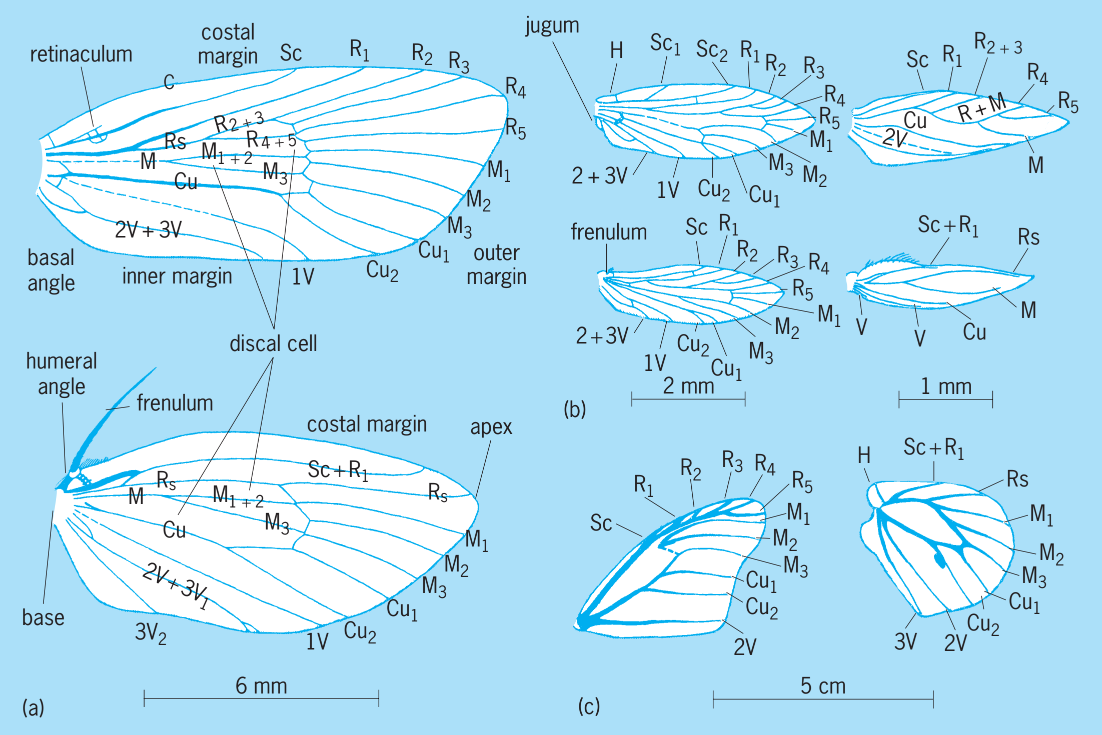


Fig. 4. Wing venation patterns with the veins labeled. (*a*) Male *Acrolophus popeanellus* (Tineidae). (*b*) *Epimartyria* (Micropterygidae) and *Nepticula nyssaefoliella* (Nepticulidae). (*c*) *Danaus plexippus* (Nymphalidae). C = costa; Sc = subcosta; R = radius; R_s = radial sector; M = media; Cu = cubitus; 1V, 2V, 3V = vannal or anal veins; H = humeral vein; subscripts refer to branches (for example, R_2 is second branch of radius).

most primitive moths, such as the Micropterygoidea and Eriocranioidea, have free (exarate) pupae in which the appendages of the head and thorax are not sheathed, the segments of the abdomen are movable, and the outer covering (integument) is soft. These pupae are also decticous, having functional mandibles. All Lepidoptera with obtect pupae, including those that are incomplete, are adecticous, that is, lacking functional mandibles.

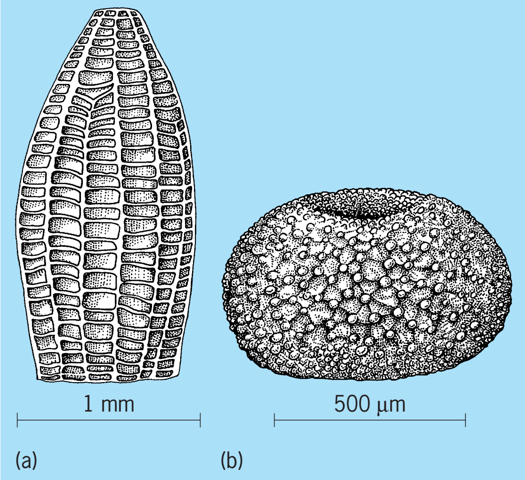


Fig. 5. Butterfly eggs. (*a*) *Satyrium acadica* (Lycaenidae). (*b*) *Anthocharis midea* (Pieridae).

Classification

The Lepidoptera were long divided into two suborders, the Rhopalocera, with clubbed antennae, notably the butterflies and skippers, and the Heterocera, with antennae of other forms, as in most moths. The butterflies and skippers form a major natural unit with over 15,000 species, still often referred to as the Rhopalocera. They are, however, closely related to various higher moths, and the simple division of Lepidoptera into Rhopalocera (butterflies) and Heterocera (moths) is not tenable.

The divisions Macrolepidoptera and Microlepidoptera often appear in older literature. The microlepidoptera comprise the more ancient groups, the most primitive of which have adults with chewing mouthparts and two pairs of similar wings (homoneurous moths), and include all of the families from Micropterygidae to Thyrididae listed in **Table 1**. Generally these are small insects, some with wingspans as small as 3 mm (0.1 in.). The macrolepidoptera include most of the larger moths and the butterflies, with wingspans typically in the 20–80-mm (0.8–3.2-in.) range, with a few reaching as much as 280 mm (11 in.). However, a few microlepidoptera, such as certain cossids and hepialids [wingspans up to 200 mm (7.9 in.) and more], are much larger than many macrolepidoptera, which can have wingspans as small as 8 mm (0.3 in.). In terms of common descent, the Macrolepidoptera do seem to form a natural group, but the microleps definitely do not, and the term Microlepidoptera is not used now, except informally (for example, micromoth).

Recent comparative research on Lepidoptera indicates that the major groups evolved in a sequential stepwise fashion. This makes natural division into just two or three large groups impossible. The very oldest moths, those with functional adult mandibles, are placed in suborder Zeugloptera, with no more than about 200 known species. Two small and very obscure groups are considered to be the next oldest known Lepidoptera, the Agathiphagidae and the Heterobathmiidae. All three of these ancient groups lack sucking mouthparts. A major innovation was evolution of the proboscis, giving rise to the so-called Glossata suborder. The living moths that represent the earliest stage of this development are the Eriocraniidae. A further novelty was development of hollow scales, a feature also shared by the great majority of living Lepidoptera. Some of these Coelolepida then evolved muscles that coil the proboscis, and the obtect type of pupa. All Lepidoptera with coilable proboscides, obtect pupae, and larvae with crotchet-bearing abdominal prolegs belong to the Neolepidoptera. When this stage was reached, these Lepidoptera evidently still had two very similar pairs of wings (Homoneura), as seen in ghost moths (Hepialidae). A further step was the evolution of a marked difference between the fore- and hindwings (Heteroneura). A subgroup of these moths then underwent a major change in the female reproductive system, developing separate orifices for copulation (vulva) and egg laying (oviduct) that are connected by a seminal duct (Ditrysia). By this stage it would seem that the Lepidoptera were increasingly successful, evolving step by step many new families. The Apoditrysia evolved with a modification of the thorax; then fusion of the first four abdominal segments of the obtect pupa led to the Obtectomera (although validity of the group based on this character is open to question, as this condition also occurs in some lower Ditrysia, including the Yponomeutoidea, Gelechioidea, and Alucitoidea). Finally, evolution of the Macrolepidoptera occurred, in which all of the many species share a special type of wing base. The familiar butterflies are subordinate members of the Macrolepidoptera, nested inside this huge and highly diverse "crown" group.

It is evident that Lepidoptera classification cannot simply reflect a set of major alternatives, such as butterflies versus moths, or Macrolepidoptera versus Microlepidoptera, but needs to mirror the successive stepwise evolution of one group after another. Thus, not only are the butterflies Macrolepidoptera,

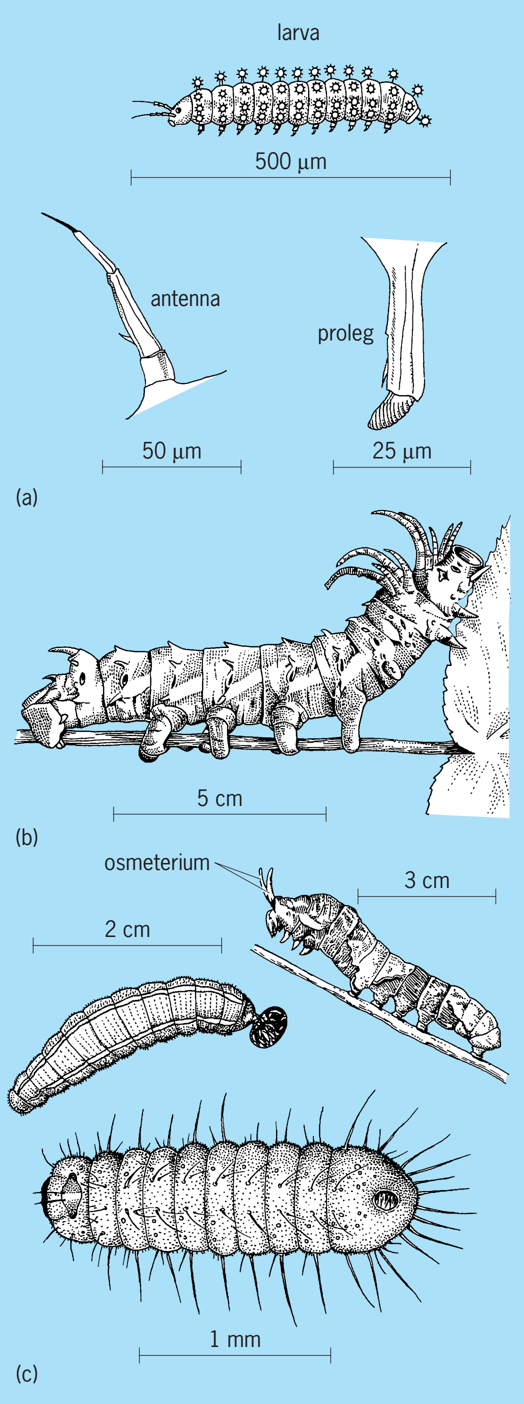


Fig. 6. Larvae. (*a*) Newly hatched larva of *Micropterix calthella* (Micropterygidae). (*b*) Hickory horned devil, larva of *Citheronia regalis* (Saturniidae). (*c*) Skipper and butterfly larvae: mature *Thorybes pylades* (Hesperiidae), showing neck; mature *Papilio cresphontes* (Papilionidae), osmeterium extruded; first instar *Satyrium liparops* (Lycaenidae), about 2 mm (0.08 in.) long.

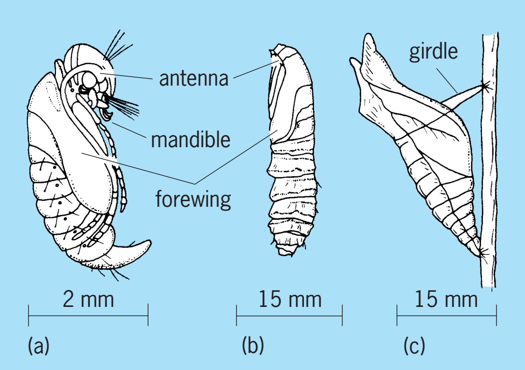


Fig. 7. Pupae. (*a*) *Dyseriocrania auricyanea* (Eriocraniidae). (*b*) *Sthenopis thule* (Hepialidae). (*c*) *Papilio troilus* (Papilionidae).

TABLE 1. Size, distribution, and common names of some families of Lepidoptera

Classification	Common name	Distribution	Number of species*
ZEUGLOPTERA			
Micropterygidae	Micropterygids	Cosmopolitan	3 (200)
OTHER NON-GLOSSATA			
Agathiphagidae	Agathiphagids	Queensland, eastern Pacific	(2)
Heterobathmiidae	Heterobathmiids	Southern South America	(10)
GLOSSATA			
Eriocraniidae	Eriocraniids	Holarctic	5 (24)
COELOLEPIDA			
Acanthopteroctetidae		North America, Peru, Crimea	4 (6)
Lophocoronidae	Lophocoronids	Australia	(6)
Neopseustidae	Neopseustids	India to Taiwan, southern South America	(10)
NEOLEPIDOPTERA			
Mnesarchaeidae	Mnesarchaeids	New Zealand	(14)
Hepialidae	Swift or ghost moths	Cosmopolitan	18 (500)
HETERONEURA			
Nepticulidae	Serpentine leaf miners	Cosmopolitan	75 (800)
Incurvariidae	Incurvariids	Mainly Palearctic and Australia	3 (100)
Prodoxidae	Yucca moths and relatives	North America (one species in southern South America)	75
DITRYSIA			
Tineidae	Clothes moths and relatives	Cosmopolitan	135 (3000)
Psychidae	Bagworms	Cosmopolitan	25 (1000)
Gracillariidae	Gracillariids	Cosmopolitan	235 (2000)
Yponomeutidae	Ermine moths	Cosmopolitan	65 (600)
Heliodinidae	Heliodinids	Cosmopolitan	21 (60)
Oecophoridae	Oecophorids	Cosmopolitan but largely Australian	225 (3200)
Coleophoridae	Case bearers	Cosmopolitan	110 (1400)
Gelechiidae	Gelechiids	Cosmopolitan	590 (4600)
Cosmopterigidae	Cosmopterigids	Cosmopolitan	(1650)
APODITRYSIA			
Megalopygidae	Flannel moths	New World, mainly tropical	11 (250)
Limacodidae	Slug moths	Cosmopolitan	50 (1000)
Zygaenidae	Foresters and burnets	Worldwide except New Zealand	20 (1000+)
Cossidae	Goat or carpenter moths	Cosmopolitan	45 (670)
Sesiidae	Clearwing moths	Cosmopolitan	120 (1200)
Castniidae	Castniids	Neotropical and Indo-Australian	(~150)
Tortricidae	Tortricids	Cosmopolitan	925 (5000+)
Alucitidae	Many-plume moths	Cosmopolitan	1 (135)
Pterophoridae	Plume moths	Cosmopolitan	130 (1000)
OBTECTOMERA			
Pyralidae	Pyralids	Cosmopolitan	500 (6,000+)
Crambidae	Snout moths	Cosmopolitan	700 (11,500+)
Thyrididae	Window-winged moths	Tropical	10 (760)
MACROLEPIDOPTERA			
Lasiocampidae	Tent caterpillars, lappet moths	Cosmopolitan except New Zealand; mainly tropical	30 (1500)
Bombycidae	Silkworm and allies	Tropical	1 (introduced) (350)
Saturniidae	Giant silkworms	Cosmopolitan	60 (1480)
Sphingidae	Sphinx, hawk, or hummingbird moths	Cosmopolitan	106 (1250)
Notodontidae	Prominents, puss moths	Cosmopolitan except New Zealand and Pacific	120 (2800+)
Noctuidae	Noctuids, owlets, underwings, millers	Cosmopolitan	2700 (35,000+)
Lymantriidae	Tussock moths	Largely African and Indo-Malayan, but with important Holarctic species	27 (2500+)
Arctiidae	Tiger moths	Cosmopolitan	200 (~6000)
Drepanidae	Hooktips	Holarctic	6 (650)
Geometridae	Measuring worms, loopers, cankerworms, carpets, waves, and pugs	Cosmopolitan	1200 (21,000)
Uraniidae	Uraniids	Tropical	(700)
Hedylidae	Hedylids	Neotropical	(40)
RHOPALOCERA			
Hesperiidae	Skippers, agave worms	Cosmopolitan	240 (3500)
Papilionidae	Swallowtails, bird-wings, parnassians	Cosmopolitan	27 (600)
Pieridae	Whites, sulfurs, orange-tips	Cosmopolitan	61 (1000)
Lycaenidae	Blues, coppers, hairstreaks	Cosmopolitan	138 (5000)
Riodinidae	Metal marks	Mainly Neotropical, but all major regions	1 (1250)
Nymphalidae	Brush-footed butterflies	Cosmopolitan	211 (6000)

*The first figure is the number of described species in North America north of Mexico. In most cases this figure is reasonably accurate. The figure in parentheses is a good estimate of the number of described species in the world. However, it is difficult to derive the total number of living Lepidoptera species from these figures. While in groups such as the Papilionidae more than 90% of existing species have almost certainly been recognized, in others, such as many families of Microlepidoptera, the figure may be well under 25%. Classification is based largely on N. P. Kristensen (ed.), 1999.

but at successively higher levels they are also members of the Obtectomera, Apoditrysia, Ditrysia, Heteroneura, Neolepidoptera, Coelolepida, and Glossata (Table 1). Butterflies are in reality just one of several advanced group of moths. Of all the major formed groups indicated in Table 1, the only group that the butterflies do not belong to is the suborder Zeugloptera, and the same is true for all the other groups listed, insofar as they are included within all the major divisions above their position in Table 1. Some of the most important Lepidoptera families are also listed in this table.

Homoneurous moths. In Table 1, the first nine families listed, up to the Heteroneura, have a primitive homoneurous wing condition, with the fore- and hindwings similar in shape and venation. They do not form a natural group, but it is convenient to deal with them together. Their fore- and hindwings are connected by a jugum and sometimes also by a frenulum. Mouthparts either are mandibulate, with the mandibles functional, vestigial, or absent (**Fig. 8**), or consist of a rudimentary proboscis. The females have a single genital opening, except in the Mnesarchaeoidea and Hepialoidea, which have two openings, one for copulation and the second for egg laying, but they lack the connecting seminal duct found in the Ditrysia. The pupae are exarate (free) or incomplete.

These primitive moths, making up less than 1% of Lepidoptera species, share a number of characteristics with the Trichoptera (caddis flies), to which the Lepidoptera are very closely related. In addition to the three groups described below, six other families of inconspicuous homoneurous moths are listed in Table 1, several of which were discovered only in the 1970s.

Superfamily Micropterygoidea. These moths make up one family, the Micropterygidae. They are small moths [up to 15 mm (0.6 in.) wingspan but usually much less], and the adults have asymmetrically toothed, functional mandibles and lack even a rudimentary proboscis. The galea is short and the adults feed on pollen or fern spores. The larvae, which feed variously on mosses, liverworts, angiosperm leaves, and possibly even fungal hyphae, mostly lack abdominal appendages, although some are very unusual in having eight pairs of simple abdominal prolegs.

It has been suggested that these moths are terrestrial Trichoptera or that they should be placed in a separate order, the Zeugloptera. However, characters of the wing venation, the broad well-developed scales with numerous ridges, and many other features indicate that these insects do belong to the Lepidoptera.

Superfamily Eriocranioidea. In this group of small moths [never more than 15 mm (0.6 in.) wingspan], the adult mandibles are greatly reduced and untoothed, and the galeae form a simple and abbreviated proboscis that cannot be coiled. These moths appear to represent the earliest stage in the evolution of the Glossata. Formerly, three or even more families were placed in this group, but the Neopseustidae, Mnesarchaeidae, and Acanthopteroctetidae are now considered to represent separate, further stages in glossatan evolution (Table 1). The single remaining family, the Eriocraniidae, have legless leaf-mining larvae. The adults fly in sunshine and drink sap from injured leaf buds. The females have a cutting (piercing) ovipositor used to insert their eggs into leaf buds.

Superfamily Hepialoidea. The hepialoids are medium-sized to very large moths possessing short, not

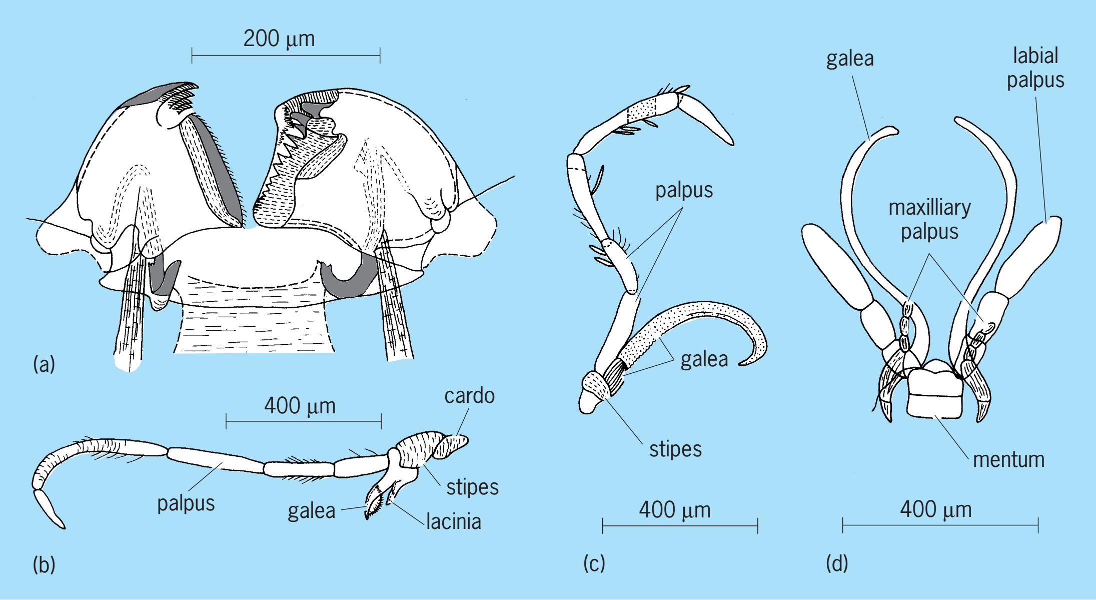


Fig. 8. Mouthparts of adult homoneurous moths. (*a*) Mandibles of *Sabatinca incongruella* (Micropterygidae). (*b*) Maxilla of *Micropteryx aruncella* (Micropterygidae). (*c*) Maxilla of *Eriocrania semipurpurella* (Eriocraniidae). (*d*) Maxillae and labium of *Mnesarchaea paracosma* (Mnesarchaeidae).

fully coilable proboscides that are probably nonfunctional. Together with the Mnesarchaeidae (Mnesarchaeoidea), they form the Exoporia, a group representing the first stage in the evolution of the Neolepidoptera. Hepialoid larvae are borers in stems or roots, either feeding in the tunnels they create or emerging to eat vegetation; some feed on decaying wood and fungi. The adults are rapid flyers and mostly crepuscular (active during twilight or preceding dawn), giving the common name swift moths or ghost moths. There are five families, the best known of which is the Hepialidae with about 500 known species. Some have wingspans approaching 250 mm (9.8 in.).

Heteroneura. Fore- and hindwings in the Heteroneura are markedly different in shape and venation. Usually they are connected by a frenulum and retinaculum. The sucking mouthparts are occasionally vestigial, but the majority of heteroneurans are able to feed on nectar, juices from rotting fruits, and other liquids. The females of all except the Nepticuloidea and Incurvarioidea have two genital openings. The pupae are obtect. The Heteroneura appear to be a major natural group that comprise all the rest of the Lepidoptera, including the Macrolepidoptera. This section deals with the heteroneurous microlepidoptera, leaving the macrolepidoptera to the third section.

Superfamily Nepticuloidea. Two very closely related families are included, Nepticulidae and Opostegidae. These tiny moths have wing spines and females have a single genital opening, but they differ from the Incurvarioidea in having reduced venation, a large eye-cap at the base of the antenna, and a short nonpiercing ovipositor. The larvae, with the exception of some gall-making species (insect galls are enlargements or swellings of plant tissue due to the feeding stimulus of the insects) of the genus *Ectoedemia*, are miners in leaves, bark, and rarely fruits. The pupae have the leg bases free. Many species of the genus *Nepticula* have a wing expanse in the 3–5-mm (0.1–0.2-in.) range, being the smallest insects in the order.

Superfamily Incurvarioidea. Six families are included in this group, with all members having an extensible, piercing ovipositor in the female, and a portable case in the last larval instar. As in the Nepticuloidea and homoneurous moths, the females have a single genital opening. The Incurvariidae have wings covered with microscopic spines, and almost complete venation. The larvae are seed, leaf, stem, or needle miners. In the subfamily Incurvariinae, the larva is first a miner and then a case bearer. With up to 300 species, the fairy moths (Adelidae) are the largest incurvarioid family. Male adelids have spectacularly long antennae, and some species make swarming flights in bright sunshine; the larvae feed on dead leaves and low plants.

The yucca moth, *Tegeticula yuccasella*, belongs to the family Prodoxidae, almost all members of which are North American. This small white moth has an obligatory relationship with yuccas. The females gather pollen with their specialized mouthparts and fertilize the yucca flower. The eggs are laid, using the piercing ovipositor, in the plant ovary, where the larvae grow by eating some of the developing seeds.

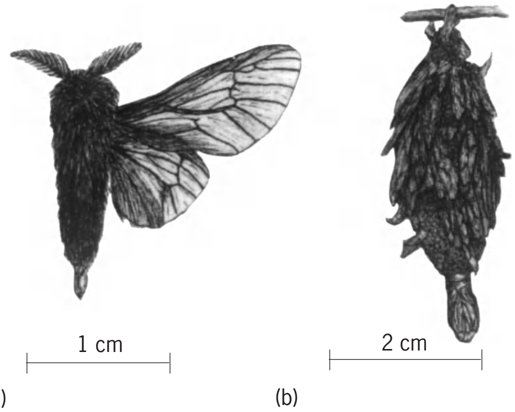


Fig. 9. ***Thyridopteryx ephemeraeformis*** **(Psychidae). (*a*) Male. (*b*) Male case.**

Superfamily Tineoidea. The Tineoidea formerly included a large and heterogeneous assemblage of moths, but only four tineoid families are now accepted, with the others groups variously reassigned. The tineoids can be regarded as the oldest or most primitive members of the Ditrysia.

The Psychidae comprise the remarkable bagworms. The males are hairy, strong-bodied, swift-flying moths with reduced mouthparts. They are relatively large, some with a wing expanse of about 25 mm (1 in.). The females are wingless, legless, and often sluglike. They live concealed in bag-shaped cases made by the caterpillars. The best-known North American representative is *Thyridopteryx ephemeraeformis* (**Fig. 9**). In this species, the larva fastens the bag to a twig and pupates within it. The vermiform (wormlike) female emerges from the pupa and moves to the bottom of the bag, where she is fertilized. This is accomplished from the outside by the highly specialized extrusible genitalia of the male. The female deposits her eggs in the bag, then drops to the ground and dies.

The Tineidae are a family of small, mostly drab moths that often have an erect tuft of yellowish scales between their antennae. The proboscis is reduced or vestigial, but the wing venation is fully developed and only rarely somewhat reduced; the ovipositor is generally long and extensible. The larvae do not feed on green plants but on a variety of substrates, including detritus, fungi, lichens, and notably keratin (wool, skin, etc.). The larvae of many species are case bearers. The best-known species are the clothes moths, which include the case-making clothes moth (*Tinea pellionella*), the webbing clothes moth (*Tineola bisselliella*), and the carpet moth (*Trichophaga tapetzella*); all are important pests whose larvae devour wool and other animal products. These three species have wing expanses of less than 25 mm (1 in.).

Superfamily Gracillarioidea. This group includes four ditrysian families that share a number of specialized features. The Gracillariidae, by far the largest of the four in terms of included species, have lanceolate (shaped like the head of a lance) and widely fringed pairs of wings, well-developed proboscides, and nonextensible ovipositors; and they comprise the principal group of leaf-mining Lepidoptera. The young larvae are flattened and have bladelike mandibles with which they slash the cells of the leaf, sucking up the exuding juices. When full-grown, the larvae of some species are quite different, being normal in appearance and feeding on leaf parenchyma, either in a mine or externally. In *Phyllocnistis* the final instar does not feed, lacks legs and mouthparts (other than the spinneret), and pupates within its mine (a gallery it creates within the leaf tissue).

Superfamily Yponomeutoidea. This is an assemblage of rather primitive micromoths, with eight families somewhat uncertainly included. The Yponomeutidae represent a heterogeneous assortment of about 600 small, often brightly colored moths that lack ocelli. Most of the 80 or so species of *Yponomeuta* (ermine moths) have larvae that live in communal webs sometimes covering whole trees or shrubs. The Heliodinidae are a small family of tiny moths with ocelli and smooth heads that often have brilliant metallic spots on the forewings. The larvae are leaf miners, borers in stems, buds, or galls, or external leaf feeders; sometimes in communal webs.

Superfamily Gelechioidea. This is a very large group of some 15 families, representing a combined total of over 16,000 described species of micromoths, with wingspans in the range 4–70 mm (0.16–2.7 in.). The superfamily is represented worldwide, but certain families or subfamilies are mainly associated with particular zoogeographical regions. It is anticipated that many thousands more species of these moths await discovery. The group is characterized by the overlapping scales found on the dorsal surface of the proboscis, although this feature does recur in some Ditrysia. *See* ZOOGEOGRAPHY.

Oecophoridae are small to moderately small moths, some of which have a comb of bristles, the pecten, on the antennal scape (shaft). The larvae are varied in habits: some feed in webs or rolled leaves, whereas others are scavengers, notably in Australia where certain species specialize on the tough fallen leaves of *Eucalyptus*. *Stathmopoda* larvae feed on coccids, and some related species feed on spider eggs. *Hofmannophila pseudospretella* is an important pest of stored products, including grain, dried fruits, and furs.

Coleophoridae are small, narrow-winged moths. The young larvae of most species are leaf miners, but from the second instar they become case bearers, carrying shells made out of silk and bits of leaves, to which they add as they grow and within which they eventually pupate. Other coleophorids feed in seed heads, mine pine needles, or are inquilines in galls. *Holcocera pulverea* (subfamily Blastobasinae) is an important predator of the lac insect in India.

The Gelechiidae are a very large family of small to minute moths, usually with rounded forewings or, rarely, pointed ones. The hindwings are trapezoidal and often pointed, with the outer margin excavated or sinuate below the apex. The larvae include seed feeders, miners, borers, gall makers, and foliage feeders, and are known to attack over 80 different plant families. The group includes a number of economically important species, such as the *Angoumois* grain moth (*Sitotroga cerealella*), which infests grain both in the field and in storage, and the pink bollworm (*Pectinophora gossypiella*), an extremely important pest of cotton worldwide. The Cosmopterigidae, with over 1600 species, are the closest relatives of the Gelechiidae. While most are plant or detritus feeders, the larvae of *Euclemensia* are parasitoids, feeding internally in armored scale insects.

Superfamily Zygaenoidea. Currently classified as the first group of the Apoditrysia, the zygaenoids comprise a dozen families apparently linked by having a retractile larval head, and the second abdominal spiracle of the pupa covered by the wings. These moderately small to sometimes quite large moths have complete venation and in some cases a rudimentary proboscis; the wings are usually broad with short fringes. The larvae are stout and sluglike, and are exposed feeders.

The Himantopteridae are a curious group of about 40 species restricted to the Old World tropics. The hindwings of these small or middle-sized moths are very narrow with long ribbonlike tails, and the body and wings are covered with long hairs. The larvae of some species feed on trees (dipterocarps and oaks), but it has been suggested that others live in termite colonies. Supposedly the newly emerged moths escape when attacked by the termites, sacrificing their long hairs that easily pull free and their expendable tails. Recent work has failed to substantiate this, but the potentially fascinating biology of these moths is poorly known.

The Megalopygidae are restricted to the New World and are mainly tropical. The head and thorax are densely scaled, and the abdomen is very hairy. The larvae of some species are pests of palms, ferns, and various other plants, and their penetrating setae (they are often covered by long hairs) can cause dermatitis in humans. The family is considered to be related to the Limacodidae.

The Limacodidae are a family of heavy-bodied, often brightly colored, hairy moths (**Fig. 10**). The

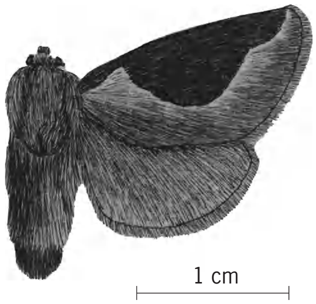


Fig. 10. ***Prolimacodes badia*** **(Limacodidae).**

larvae are short and especially sluglike (slug caterpillars), with a large head concealed by the thorax, and furnished with various spines, tubercles, and even gelatinous warts. The thoracic legs are small or minute, and the abdominal prolegs are replaced by midventral suckers that are used, together with a fluid secreted by the spinneret, to stick the caterpillars tightly to foliage. The best-known North American form is the saddleback caterpillar (*Sabine stimulea*) which, like many limacodid larvae, has irritating (urticating) hairs.

The Zygaenidae are medium-sized, often brightly colored, mainly day-flying moths found in most regions (but not New Zealand), with over 1000 known species. They appear to be chemically protected, and if attacked many produce cyanide, a poison to which they appear to be resistant. Many zygaenids are involved in mimetic associations, notably the Oriental subfamily Chalcosiinae, some of which look very similar to aposematic (pertaining to colors of an organism that give warning of its special means of defense against enemies) butterflies such as Danainae. The larvae are rather stout and feed openly on plants. In some groups, notably the large genus *Zygaena*, the pupa is formed within a characteristic parchmentlike cocoon spun on a grass stem or similar substrate. *See* PROTECTIVE COLORATION.

Superfamily Cossoidea. Two families are included. Because the larvae bore tunnels in wood and smell rather strongly, the Cossidae are commonly called carpenter moths or goat moths. They are heavy-bodied insects, sometimes very large [maximum 235-mm (9.2 in.) wingspan], with the abdomen extending well beyond the hindwings. The mouthparts are largely rudimentary. The median vein (M, Fig. 4) is present and usually forked within the discal cell (a large cell in the central, or disc, part of the wing) of both wings. Although the larvae of some species feed externally on stems or roots, most are tunnelers, even in hardwood trunks. *Prionoxystus robiniae*, the best-known American cossid, is very destructive to a large variety of deciduous trees, and *Zeuzera coffeae* damages numerous species, including coffee. Like the Cossidae, the Dudgeonidae, a group of six stem borers found in the Old World tropics characterized by an abdominal hearing organ, also have a very strongly reduced proboscis.

Superfamily Sesioidea. This group of three families is thought to be very closely related to the Cossoidea. The Sesiidae are called clearwing moths because of the large, transparent, scaleless areas on the wings. Veins Sc and R_s on the hindwing appear to be absent, but are actually concealed in a unique costal fold that acts as an additional wing-coupling mechanism. Diurnal, with contrasting black and transparent patterns, or more brightly colored, many of these moths are excellent wasp mimics. The larvae are mostly borers, and many are economic pests, including the currant borer (*Synanthedon tipuliformis*), peachtree borer (*Sanninoidea exitiosa*), and squash-vine borer (*Melittia satyriniformis*). The South American *Synanthedon coccidivora* is carnivorous on coccids.

The Castniidae are large [wingspan range of 24–190 mm (0.9–7.5 in.)], diurnal, butterflylike moths with broad wings, clubbed antennae, and fusiform eggs. The larvae bore into monocots, and some are minor pests. A proboscis may be present or (rarely) vestigial. These moths have been considered to be related to the butterflies, but the resemblances must be due to convergence, and some are mimetic of butterflies and other moths. They are found in the Neotropical and Indo-Australian regions.

Superfamily Tortricoidea. These small, broad-winged moths comprise just a single very large family (over 5000 species), the Tortricidae, currently divided into three subfamilies. The best character for the group as a whole appears to be the large, flat ovipositor lobes of the females.

The hindwings of the moths belonging to subfamily Olethreutinae usually have a fringe of long hairs on the upper side, running along the basal part of the cubitus. The larvae are generally hidden feeders and live in rolled leaves, in foliage webbed together, or inside fruits. The subfamily contains a number of agriculturally very undesirable species, notably the codling moth (*Cydia pomonella*), a serious pest of apples and other fruits. The genus *Cydia* (formerly called *Laspeyresia*) also contains the curious Mexican-jumping-bean moth, *C. saltitans*. The violent movements of the larvae feeding inside are responsible for the jumping action of the beans.

Moths placed in the subfamily Tortricinae generally lack the long cubital fringe on the hindwing characteristic of the Olethreutinae. The spruce budworm (*Choristoneura fumiferana*), which belongs to the large tortricine tribe Archipini, is probably the most injurious tortricid. In many places, especially eastern Canada, the larvae of this moth have defoliated vast areas of coniferous forest. Members of the third and smallest subfamily, the Chlidanotinae, are largely tropical and relatively poorly known.

Superfamily Alucitoidea. Two small families of rather delicate micromoths are included. The pupae have immovable abdominal segments I–IV. The Alucitidae are the larger of the two families, with larvae that bore into buds, shoots, flowers, and fruits or make galls; the African *Alucita coffeina* is a pest of coffee. The adults of most species have the forewings divided into six featherlike plumes, and the hindwings into six or seven. Although this superfamily is now believed to be very closely related to the plume moths (Pterophoridae), the striking similarity of wing division appears to be a case of convergence.

Superfamily Pterophoroidea. The Pterophoridae are known as the plume moths. The wings of most species are divided into featherlike plumes, of which there are usually two in the forewing and three in the hindwing (**Fig. 11**), similar to the Alucitidae. Although the alucitoids are now thought to be closely related, the evidence is ambiguous; the Pterophoroidea may be closer to the Pyraloidea. The moths lack ocelli and have characteristic slender bodies and long legs, with the wings at rest held outstretched; segments I–IV of the pupae are movable. The larvae feed exposed or are borers. *Buckleria* is remarkable for feeding on insectivorous sundews

Fig. 11. *Platyptilia carduidactyla* (Pterophoridae).

(Droseraceae); the North American *Platyptilla carduidactyla* (Fig. 11) is a minor pest of artichokes.

Superfamily Pyraloidea. The Pyraloidea comprise one of the largest family-level divisions of the Lepidoptera, with over 16,000 named species, and the expectation is that this will rise to 30,000 or more when the taxonomy is worked out. Currently the group is divided into two families: the Pyralidae with 5 subfamilies and Crambidae with 15. All are linked by possession of a unique type of hearing organ located on the first abdominal segment. They are small to quite large moths, with a wing expanse in the range of 10–150 mm (0.4–6 in.), usually about 20–35 mm (0.8–1.4 in.). The hindwings often have three anal veins, while the legs are usually long and slender.

The Pyralidae differ from the Crambidae in having the case of the hearing organ almost completely closed. The relatively small subfamily Galleriinae includes the bee moth, or wax worm (*Galleria mellonella*), which lives in beehives. The larvae feed on the wax at night and destroy the combs; the species occurs throughout the range of the honeybee. Subfamily Pyralinae reaches its richest development in the Old World tropics. The meal moth (*Pyralis farinalis*) is a cosmopolitan pest of stored products. The Phycitinae comprise a large group of moths (almost 4000 species) in which the frenulum of the female, like that of the male, is a simple spine rather than a bundle of bristles. The larvae have very diverse habits, being leaf rollers, case bearers, borers, and stored-products pests, and some (genus *Laetilia*) are predacious on coccids. The Indian meal moth (*Plodia interpunctella*) is cosmopolitan, feeding on a wide variety of stored products, especially cereals. Another extremely important pest is the Mediterranean flour moth (*Anagasta kuehniella*), which infests cereals throughout the world. In contrast to these harmful species, the subfamily Phycitinae also contains *Cactoblastis cactorum*, imported from South America into Australia, Hawaii, South Africa, and other areas to help control introduced *Opuntia cactus*, which had ruined millions of acres of pasture. This successful program is an outstanding example of biological control. *See* ENTOMOLOGY, ECONOMIC; INSECT CONTROL, BIOLOGICAL.

Crambidae are separated from the Pyralidae by the wide aperture to the hearing organ case and differences in the male genitalia. Subfamily Crambiinae (snout moths) contains small insects common in marshes and grasslands. In this group the labial palpi (Fig. 1*a*) are long and porrect (extended forward), giving the adults a beaked appearance. Species of the small subfamily Schoenobiinae bore in the stems of marsh-living grasses, with *Scirpophaga* and *Rupela* being important rice pests. Members of the subfamily Nymphulinae are notable for being almost entirely aquatic, and in some species the larvae are able to swim from plant to plant. The larvae of some nymphulines have tracheal gills, and the pupae are usually enclosed in a cocoon below the surface, from which the adults emerge through the water. In the genus *Acentria*, most female adults are wingless and reproduce parthenogenetically, never emerging from the water; males and winged sexual females emerge periodically, swarm, mate, and disperse. With about 7400 species, the Pyraustinae are the largest group of pyraloids and include some relatively large moths with wingspans of 30 mm (1.2 in.) or more. This subfamily includes the infamous European corn borer (*Ostrinia nubilalis*), the grape leaf folder (*Desmia funeralis*), and numerous other economically damaging species.

Superfamily Thyridoidea. This mainly tropical group comprises a single family, the Thyrididae, with about 760 described species and many undescribed. They have often been included in the Pyraloidea, but they lack abdominal hearing organs and have a peculiar resting posture in which the body is raised up steeply and the wings held outstretched. Larvae of the Oriental genus *Glanycus* can stridulate (make audible sounds, produced by rubbing various body parts together), possibly a warning signal as these brightly colored moths appear to be chemically protected.

Macrolepidoptera. All of the remaining groups dealt with below are macroleps. Unlike the microlepidoptera, there is some morphological evidence that the macros form a natural group, and they are now formally recognized as such. The Macrolepidoptera are the crown group of the order and include some 100,000 different sorts—more than half of all known Lepidoptera species. The largest moths and butterflies are included, and even the smallest are larger than nepticulids and other micros. Because of their size, the larvae of most macros are external feeders, but there are exceptions.

Superfamily Lasiocampoidea. Two families, somewhat uncertainly united but both apparently related to the Bombycoidea, are included in this group. The Anthelidae are a small family restricted to the Australian region. Male Lasiocampidae always have bipectinate antennae (Fig. 3*b*), right to the tip; the females are usually similar but with shorter rami. This widespread, largely tropical family contains the eggar and lappet moths, and the familiar tent caterpillars of the genus *Malacosoma*. In this genus the larvae live together in large silken nests. Those of *M. americana* are a common sight on trees and shrubs in the eastern United States.

Superfamily Bombycoidea. A group of nine families of often stout-bodied large moths, including some of the most spectacular insects in the whole order. These families appear to be reliably grouped together by a number of characters, including the anterior fusion of the forecoxae in the last instar larvae, and a peculiarity of the forewing radial venation.

The family Eupterotidae represent a group of about 300 species, characterized by the last tarsal segment of the metathoracic leg of the female having a midventral row of spines. On this basis, the sexually dimorphic (having secondary morphological differences between the sexes) African genus *Hibrilides*, in which the females apparently mimic aposematic butterflies, has recently been included in this largely tropical family. *See* SEXUAL DIMORPHISM.

The Bombycidae include the commercial silkworm (*Bombyx mori*). The larvae of this totally domesticated species feed on mulberry leaves, and resemble hawkmoths in that they possess a small caudal horn. The silk from the white or yellow cocoons has been used since prehistoric times. The caterpillars are subject to the ravages of a disease known as pébrine, caused by the protozoon *Nosema bombycis*. They are also attacked by two viral diseases known as grasserie and flacherie. The Bombycidae are recognized as a group based on characters of the thorax and pupa, and worldwide the family includes about 350 species. *See* MICROSPORIDEA.

Saturniidae are the giant silkworms, medium-sized to extremely large insects having a single, often weak humeral vein in the hindwing. The mouthparts are reduced or vestigial, and the antennae are strongly bipectinate. As in the Lasiocampidae, the rami of the male antennae are conspicuously longer than those of the female. The frenulum is absent. The larvae usually bear spiny processes known as scoli. Pupation takes place in large silken cocoons, which are suspended from twigs or formed in leaf litter on the ground. Females of species *Attacus atlas* of the Himalayan foothills and *Coscinoscera hercules* of Australia have wingspans of up to 250 mm (9.8 in.) and possibly more, and these moths are thought to have the largest wing areas in the Insecta. A number of saturniids produce usable silk, which has long been gathered in the Orient. Among these are the Japanese oak silkworm (*Antheraea yamamai*) and the muga silkworm (*A. assamensis*).

The nearly 1500 known species of Saturniidae, which are mostly tropical, are currently divided among nine subfamilies. In North America, the best-known species of the subfamily Saturniinae are the cecropia (*Hyalophera cecropia*), polyphemus (*Antheraea polyphemus*), luna (*Actias luna*), promethea (*Callosamia promethea*), and cynthia (*Samia cynthia*). Likewise in North America, the regal moth, or hickory horned-devil (*Citheronia regalis*), and the imperial moth (*Eacles imperialis*) are the most familiar members of the subfamily Ceratocampinae. In subfamily Hemileucinae, the buck moths and relatives, the io moth (*Automeris io*) is noteworthy because its common larvae, like those of some other members of the subfamily, are equipped with urticating (poisonous) hairs that can cause considerable irritation to the unwary who handle them. The scoli of some tropical species release powerful anticoagulants that can induce bleeding from mucous membranes.

Sphinx, hawk, or hummingbird moths constitute the family Sphingidae, a worldwide group of some 1250 species. Morphological peculiarities of the larvae, pupae, and adults suggest that the family is a natural group. Mainly medium-sized to very large, these heavy-bodied moths have extremely rapid flight. The adults are mostly crepuscular or nocturnal, but a few genera are diurnal, including the hummingbird hawks. They have rather characteristic antennae that are thickened and have a pointed apex, while the proboscis is well developed and often extremely long. The wings are narrow, with the hindwing much shorter than the forewing. A frenulum is present, rarely reduced to a tubular thickening. The larvae are external feeders and usually have a conspicuous caudal horn. The pupa is formed in a cell in the ground or in a loose cocoon at the surface, and in some species the long proboscis is housed in a projecting case resembling a jug handle. One of the most widely distributed species is the white-lined sphinx (*Hyles lineata*). Well known also is the Old World death's-head sphinx (*Acherontia atropos*), the adults of which will enter beehives in search of honey. Of economic importance in the Western Hemisphere are the tomato hornworm (*Manduca quinquemaculatus*) and the tobacco hornworm (*M. sexta*), both of which are pests of solanaceous plants. *Erinnyis ello* is a serious pest of cassava, rubber, and ornamental euphorbs in South America, while *Cephonodes hylas* is a major pest of coffee in many parts of Africa and Asia.

Superfamily Noctuoidea. This huge superfamily contains some 50,000–70,000 species, by far the largest family-level group in the Lepidoptera, and likely to remain so even if the Gelechioidea and Pyraloidea are ever fully worked out. Although rather uniform in basic structure, which has made their classification very difficult, the noctuoids are remarkably varied in size, coloration, and biology. As measured by wingspan, they include the smallest macromoths, down to about 8 mm (0.3 in.), and the largest: the South American *Thysania agrippina* sometimes attains wing spans of 280 mm (11 in.). Currently eight families, some with numerous subfamilies, are included. All members of this great assemblage are linked by possession of a special metathoracic hearing organ that is capable of perceiving ultrasound. This organ not only may be used to detect and escape from bats but may also function in courtship signaling. *See* PHONORECEPTION.

Notodontidae are commonly called prominents, or puss moths. Together with two very small families (Oeonsandridae from Australia and the New World Doidae), they are separable from most other Noctuoidea by the apparently three-branched, or trifid, cubitus. The larvae are external feeders, and the pupa is formed in a cell in the ground or in a loose cocoon on the surface. The Nearctic *Datana* *ministra*, the yellow-necked caterpillar, and *Schizura concinna*, the red-humped caterpillar, are pests on apples and other trees.

Noctuidae (once known as the Phalaenidae) are the owlet moths, an extremely large grouping of 35,000 or more mostly dull-colored, medium-sized moths. Evidence that this family, which is currently divided into 28 subfamilies, forms a natural group is unconvincing, and significant changes to the classification should be anticipated. As in most of the remaining Noctuoidea, the noctuid cubitus is usually quadrifid, but this is not a constant feature.

As exemplified by the underwing genus *Catocala*, the forewings of the Noctuidae are almost always dully and cryptically colored. At rest, the forewings cover the hindwings, which may be strikingly colored. Most moths attracted to lights at night belong to this family. The larvae are mostly exposed foliage feeders, but a few such as *Papaipema* are borers. Some *Eublemma* species prey on scale insects, one being an important enemy of the commercially valuable lac insect. Pupation is usually in the ground. The family includes many agricultural pests: the cutworms, *Euxoa* and *Peridroma*, attack a large variety of plants, as does the army worm, *Pseudaletia unipunctata*. The former derive their name from the habit of cutting off shoots at the surface of the soil without consuming them; the latter derive their name from the fact that they often appear in vast numbers. The exceedingly important pest *Helicoverpa armigera* is variously known as the corn earworm, cotton bollworm, and tomato fruitworm.

Lymantriidae are the tussock moths. Evidence that this group of medium-sized moths form a natural group is largely lacking, but the eversible middorsal gland on abdominal segment VI of the larva may be a reliable character. The antennae of the males are broadly pectinate. The end of the female abdomen often has a tuft of detachable hairs which are deposited on top of the eggs. The hairy larvae usually have prominent toothbrush tufts. The most familiar species is the infamous European gypsy moth (*Lymantria dispar*) which, along with the destructive brown-tail moth (*Euproctis chrysorrhoea*), was imported into New England in the latter half of the nineteenth century.

Arctiidae are a relatively large family of often strikingly colored, frequently heavy-bodied insects, the tiger moths. The arctiids are characterized by a pair of eversible pheromone glands on the dorsum of the female abdomen, connected with the anal papillae. The larvae are often very hairy and feed either exposed or in webs. Larval hairs are incorporated into the silken cocoon. In North America, the best-known member of the subfamily Arctiinae is probably the banded wooly bear caterpillar (*Pyrrharctia isabella*). There is a widespread misconception that the banding pattern of this larva predicts the severity of the coming winter. In Europe, the wooly bear of the spectacular garden tiger moth (*Arctia caja*) is equally well known. The larvae of the Lithosiinae are unusual in feeding on algae, lichens, liverworts, or mosses. The third arctiid subfamily currently recognized, the Syntominae, includes many remarkable wasp mimics. Despite their great beauty and fascinating biology, the Arctiidae remain poorly known, and many discoveries can still be made.

Superfamily Drepanoidea. This relatively small superfamily includes just two families, linked by the unusual form of the larval mandibles, which have a large flat lateral area marked off beneath by a distinct ridge. The group is thought to be closely related to the Geometroidea. Epicopeiidae are a small group of brightly colored day-flying species that appear to mimic a variety of other diurnal Lepidoptera; they are known only from the Palearctic and tropical Asia. The more familiar Drepanidae include about 650 species divisible into three subfamilies, of which the Drepaninae are the largest and include the hooktip moths, named for their forewing shape. The Drepanidae are characterized by an abdominal hearing organ located between tergum I and sternum II of a type not found in any other Lepidoptera, including the Geometroidea.

Superfamily Geometroidea. Three families are included, the somewhat butterflylike Sematuridae (about 40 species, which lack abdominal hearing organs), the Uraniidae (about 700 species, also somewhat butterflylike but with abdominal hearing organs), and the Geometridae (a very large group of more than 20,000 species characterized by the unique form of the abdominal hearing organs). The three families are linked together by a character of the larval spinneret.

The tropical Uraniidae include some slender-bodied, brilliantly colored, diurnal insects that lack a frenulum and are often mistaken for butterflies. The multiple-tailed metallic Madagascan sunset moth, *Chrysiridia madagascariensis*, is considered by many to be the most beautiful lepidopteran. It also offers a fascinating example of asymmetry: almost all Lepidoptera have highly symmetrical wing patterns, but this is not the case in *Chrysiridia* (and some other uraniids) in that the black spots on the hindwing orange area, and the black markings on the forewing green areas, invariably differ right to left. Uraniid larvae are diverse in form but have a complete set of prolegs.

The Geometridae include the measuring worms, loopers, and cankerworms, which generally turn into small and medium-sized moths with slender bodies and relatively broad wings. The females are occasionally wingless (apterous). The larvae have reduced or absent anterior prolegs, and usually only those on segments VI and X are well developed. They proceed with a characteristic looping (earth-measuring) motion. The larvae are usually exposed feeders, but sometimes live in folded leaves and often bear a striking resemblance to twigs. This resemblance is enhanced by their habit of resting with the body held rigid at an angle to the branch while holding on with the terminal prolegs. Pupation takes place in the ground or in a weak silken cocoon.

This very large family is currently divided into eight subfamilies. Among the economically important Nearctic geometrids are the spring and fall cankerworms, *Paleacrita vernata* (subfamily Ennominae) and *Alsophila pometaria* (subfamily Alsophilinae), which attack a large variety of trees. The famous peppered moth (*Biston betularia*), so much studied regarding industrial melanism (below), is another member of the Ennominae. Also of interest is *Idaea bonifata* (subfamily Sterrhinae), the herbarium moth, whose larvae are pests in collections of dried plants. The beautiful emerald moths belong to the subfamily Geometrinae.

Superfamily Hedyloidea. This is a small group of about 40 mothlike species with a curious resting posture, long placed in the Geometridae but recently considered to be the likely closest relatives of the Rhopalocera (skippers and true butterflies). Their early stages are remarkably similar in some ways to those of Papilionoidea; the adults stand on only four legs like Nymphalidae; and research on their DNA confirms a close relationship. The Hedylidae comprise just one genus, *Macrosoma*, found in Central and South America.

Superfamily Hesperioidea. One rather large family, the Hesperiidae, is recognized, with about 3500 known species and probably many more unrecognized ones. The skippers, or skipper butterflies, are small to moderately large, heavy-bodied, mostly diurnal insects. The clubbed antennae are set wide apart on the head and are bent, curved, or reflexed at the tip. Forelegs are fully developed and bear a large spur (epiphysis). The forewings usually have all veins arising separately from the discal cell, and the frenulum is absent except in the male of *Euschemon*. Near the base of the hindwing, a small patch of specialized scales apparently occurs in all species. The larvae have a prominent constriction, or neck, behind the head, and often live in leaves drawn together by silk. However, the caterpillars of the giant skippers (*Megathymus* group), which are borers in yucca and agave in the Americas, lack this neck. The pupa is usually enclosed in a slight cocoon.

The name skipper refers to the typically rapid erratic flight. American species include the silver-spotted skipper (*Epargyreus clarus*) the sachem skipper (*Atalopedes campestris*). In Mexico, agave caterpillars are fried and canned for human consumption. The Australian *Euschemon rafflesia* was once placed in its own subfamily, largely because of the presence of a frenulum in the male, but it is now included in the Pyrginae, a group of some 1000 species and one of six subfamilies into which the skippers are currently divided. New molecular studies based on DNA are expected to bring many improvements to skipper classification.

Superfamily Papilionoidea. These so-called true butterflies, together with the Hesperioidea, make up the Rhopalocera. The two groups share some specialized features in common, are usually regarded as each other's closest relatives, and are together perhaps most closely related to the Hedyloidea. However, it is possible that the hedylids will eventually prove to be more closely related to one of the two Rhopalocera superfamilies. The Papilionoidea are small-to-large diurnal insects [a few are very large, as measured by wingspan, sometimes attaining 250 mm (9.8 in.)], typically with clubbed antennae that are rounded at the tip and not bent or reflexed. The forewings always have two or more stalked veins arising from the discal cell. The frenulum is always absent. The larvae have no constriction behind the head and are usually exposed feeders. The pupa is naked except in the apollos (*Parnassius*) and their relatives and some Lycaenidae, and is most often suspended head-down by caudal hooks (the cremaster) or held semi-upright by a silken girdle. About 14,000 species are currently recognized worldwide, divided among five families.

The Papilionidae are the family that includes the swallowtails and parnassians. These are the only true butterflies with a fully developed foreleg bearing an epiphysis. The hindwings have only one well-developed vannal vein, except in the anomalous Mexican *Baronia brevicornis* (subfamily Baroniinae), while the larvae always have a unique eversible forked organ on the prothorax, the osmeterium. This organ dispenses a disagreeable odor that seems to function as a defense mechanism. The pupa is typically girdled. In the boreal genus *Parnassius* (subfamily Parnassiinae) and some other members of the family, females may be found with a horny pouch, or sphragis, at the tip of the abdomen. This is secreted by the male during copulation and covers the genital opening of the impregnated female to prevent access by rival males. The commonest genera—*Papilio*, *Graphium*, *Battus*, and *Parides* (subfamily Papilioninae)—contain large and attractive species, many of which have the characteristic tails that give the family its name. The birdwing butterflies, *Ornithoptera*, *Trogonoptera*, and *Troides* (Papilioninae), are among the largest and most beautiful species. In North America, the eastern tiger swallowtail (*Papilio glaucus*) has dichromatic females, one form being black-and-yellow-striped like the male, and the other being entirely dark brown or black, apparently mimicking the protected Aristolochia swallowtail (*Battus philenor*). The larva of the orange dog (*Papilio cresphontes*) is sometimes injurious to citrus, and in the Old World there are a number of species that can be serious citrus pests, including the Oriental lime swallowtail (*Papilio demoleus*), recently introduced into the Caribbean.

The Pieridae include the familiar whites, sulfurs, and orange tips. These butterflies appear to have unique types of abdominal base and antennae. The forelegs are completely developed in both sexes but lack an epiphysis. The tarsal claws are bifid, while many species exhibit striking sexual dimorphism. However, most are basically white, yellow, or orange. The larvae feed on plants of the families Brassicaceae (mustards and cabbages), Fabaceae (beans, peas, and other legumes), and Capparidaceae (capers). The pupae are suspended by a girdle, as in the Papilionidae.

Butterflies of this family are often extremely abundant. The giant sulfurs (*Phoebis*) and others often migrate in huge swarms. The larvae of the Neotropical species *Eucheira socialis* are gregarious, living together in a nest; and in Mexico, where their larvae are harvested for food, there is a population of this butterfly that is semidomesticated. The cabbage worm or small white (*Pieris rapae*) is one of the most economically important butterflies, attacking cultivated crucifers in Europe and North America. The alfalfa butterfly (*Colias eurytheme*) is a major pest in America.

The Lycaenidae include the blues, gossamers, hairstreaks, harvesters, and coppers. This large family, with about 5000 known species and many still to be recognized, is very closely related to the Riodinidae, and the latter have often been included within the lycaenids. All of these butterflies have a unique thorax in that the lamella of the mesodiscrimen is not complete to the furca. In male Lycaenidae the prothoracic legs, which always lack tarsal claws, are functional; but in Riodinidae they are reduced and not used for walking. In both families the eyes are emarginate at the antennal bases or at least contiguous with them. The larvae are mostly flattened, and the pupa is usually girdled. The larvae often have a mutualistic relationship with ants, which tend and protect them in return for honeydew which they secrete; however, a number are predacious on ant larvae or on plant bugs attended by ants. The ant organs of the Riodinidae differ in structure from those of the Lycaenidae, and recent molecular work confirms that the two groups are separate but are each other's closest relatives.

Lycaenidae are small to medium-sized insects that are often metallic blue or green on the upper surface. Species such as *Brephidium exilis* and *Zizula hylax* (subfamily Polyommatinae) have a wingspan of about 15 mm (0.6 in.) and are among the smallest butterflies, all of which are lycaenids. The largest lycaenid is the Indo-Australian *Liphyra brassolis* (subfamily Miletinae), with a wing expanse of about 75 mm (3 in.); its extraordinary armored caterpillars live in weaver ant nests and prey on their larvae. The harvesters (*Miletus, Spalgis, Fenesica*; Miletinae) have predacious larvae that feed on homopterans attended by ants, such as aphids, membracids, and coccids. In the United States, the gray or common hairstreak, *Strymon melinus* (subfamily Theclinae), has at times been an important pest of hops, cotton, and other crops. In various parts of the world, a number of lycaenids are significant pests, notably of beans and other legumes, often feeding on flowers or boring in pods. Species making up the relatively small group of copper butterflies (subfamily Lycaeninae) do not have associations with ants. Widespread in the Holarctic and parts of Asia and Africa, they have a remarkable distribution that includes endemic species in New Zealand, with the closest other members of the group in New Guinea.

The Riodininae, or metal marks, are primarily Neotropical. They are small butterflies, often brilliantly colored. Like the Lycaenidae, many have associations with ants, but it is currently thought that the Riodininae have probably evolved independently. The extraordinary gray and semidiaphanous Peruvian species *Styx infernalis* was formerly placed in its own family or subfamily, but recent discovery of its early stages confirms that it belongs to the Euselasiinae, a group of about 150 species. The 100 or so species that make up the Nemeobiinae form the Old World representatives of the family, and include the European duke of burgundy fritillary (*Hamearis lucina*). The third subfamily, Riodininae, are exclusively American, and are by far the largest group, exhibiting an extraordinary diversity of color and form, including the beautiful *Helicopis* butterflies with six tails on the hindwings and raised underside silvery markings that look like drops of water.

The Nymphalidae are a family of four-footed butterflies, having the prothoracic legs greatly atrophied in both sexes, especially the males, and useless for walking. However, the best character for bringing all 6000 or so included species together is the tricarinate antenna, which has three parallel ridges running along their entire inner-lower surface of the antennal shaft. The female abdomen of those species examined has a von Siebold organ, a structure unknown elsewhere in the butterflies. Although subdivision of this very familiar group of butterflies has been a source of considerable disagreement over the last 200 years, advances in the application of DNA sequence data, coupled with systematic reviews of their morphology and greater knowledge of the early stages, has recently led to greater clarity and an emerging consensus.

The Libytheinae are a small subfamily of less than 20 species divided into two genera, *Libythea* and *Libytheana*. Their long, forwardly directed labial palpi have earned them the name snout butterflies. The eggs and caterpillars of these little butterflies are superficially similar to those of the Pieridae. Formerly often placed as a separate family, the snouts are now widely accepted as the first side branch of the nymphalid family tree.

The next side branch now appears to be the Danainae, a worldwide group of usually warning-colored butterflies which have an acrid taste to humans. Supposedly distasteful to most predators, they appear to be primary models in a great number of mimicry complexes. The weak-flying Neotropical glasswings were formerly placed in their own subfamily, the Ithomiinae, but they are structurally and biologically very similar to the true danaines, and are now included by some in the same subfamily. Many have broad transparent areas in the wings in which the scales are reduced to short hairs; others have bright yellow, black, and orange tiger patterns. The widely distributed and migratory monarch, or milkweed, butterfly (*Danaus plexippus*) is the best-known representative of the subfamily. In North America, it is mimicked by the viceroy (*Limenitus archippus*), a member of another nymphalid subfamily, the Limentidinae. In the Indo-Australian tropics, the largest group is the crow butterflies (*Euploea*), many species of which are black or largely so.

The very large cosmopolitan subfamily Satyrinae, containing the woodnymphs, meadow browns, graylings, and arctics, is characterized by the typically weak bouncing flight of the imagines (plural form of imago, which is the sexually mature, usually winged stage of an adult insect) and the bladderlike swellings to the basal sections of the forewing veins of most species. The larvae feed primarily on grasses and other monocots, with a few on lower plants such as lycopods. *Erebia* and *Oeneis* penetrate extremely inhospitable arctic and alpine situations, and numerous other satyrines abound in Patagonia, in the high Andes, and in many other montane regions, where many species still await discovery. Subfamily Morphinae is a much smaller but entirely tropical group of large butterflies closely allied to the satyrines. The brilliant metallic-blue species of the genus *Morpho* are among the most beautiful insects. The owl butterflies (*Caligo*) have undersides bearing large and very conspicuous eyespots, and also belong to this group. Two more subfamilies that seem to belong with the satyrine-morphine group are the Sino-Himalayan Calinaginae and the pantropical Charaxinae. This last group includes some of the fastest flying and most beautiful butterflies. Most but by no means all species in this entire assemblage feed as larvae on monocots and are strongly attracted as adults to fallen and fermenting fruit.

The crown group of the Nymphalidae comprises a further series of subfamilies, of which the Heliconiinae and Limenitidinae may form the first major branch. Well-known heliconiines include *Argynnis* and *Boloria* (temperate region fritillaries); *Vindula*, *Cethosia*, and *Acraea* in the Old World tropics; and *Heliconius*, famous for longevity and remarkable mimetic associations, in the Neotropics. The limenitidines comprise *Limenitis* (white admirals) together with numerous tropical genera and species. The final series of subfamilies, which may form a terminal group, include the Cyrestinae (tropical map butterflies), Biblidinae (eighty-eights and numerous tropical species), Apaturinae (purple emperors and hackberry butterflies), and Nymphalinae. This last subfamily includes some of the most familiar butterfly genera, such as *Polygonia* (angle-wings and commas), *Nymphalis* (tortoise-shells), *Vanessa* (red admirals and thistle butterflies), *Junonia* (buckeyes), and *Euphydryas* (checkerspots). The subfamily is also represented in the tropics by numerous genera and species, including the genus *Kallima*, or Indian leaf butterflies, the undersides of which bear a remarkable resemblance to dead leaves. *See* PROTECTIVE COLORATION.

Biological Aspects

The Lepidoptera are a group on which much research remains to be done. A great deal is still unknown about the genetics, developmental biology, physiology, and ecology of these insects despite the fact that butterflies and moths have proved useful as experimental animals in all of these fields. Economically, the larvae of many species are injurious to certain crops (**Table 2**), often causing severe losses; while a few, such as the silkworm (*Bombyx mori*), are the basis of significant industries. In recent years, trade in living tropical butterflies has grown rapidly for the butterfly house industry (insect zoos). A small number of Lepidoptera species cause medical problems, notably larvae with irritating spines and hairs, some of which can cause severe dermatitis.

Ecology and distribution. The variation in larval habits has been discussed under the different families. Lepidoptera of all stages are subject to the attacks of a large number of predators, including birds, mammals, lizards, frogs, and spiders. They also must be wary of rapacious insects, such as dragonflies, preying mantids, pentatomid bugs, asilid flies, and vespoid wasps. Certain wasps (family Sphecidae) paralyze caterpillars, lay eggs in them, and place them in specially constructed cells. Others (braconids and ichneumonids) place their eggs in the caterpillars without paralyzing them. In both cases the caterpillars serve as food for the growing wasp larvae inside them. Many chalcid wasps oviposit (lay or deposit eggs) in the eggs of Lepidoptera. *See* PREDATOR-PREY INTERACTIONS.

One mite, *Myrmonyssus phalaenodectes*, infests the hearing organs of a variety of moths; and another, *Otopheidomenis zalestes*, is restricted, so far as is known, to the genus *Zale* (family Noctuidae). Pseudoscorpions of the genera *Atemnus*, *Stenowithius*, and *Apocheridium* have been found on adults of various species. Lepidoptera are also subject to viral, bacterial, protozoan, and fungal infections. *See* INSECT PATHOLOGY.

The Lepidoptera penetrate almost every place on Earth with the exception of Antarctica. Butterflies of the genus *Boloria* have been taken at Alert on northern Ellesmere Island, about 400 mi (650 km) from the North Pole. Arctic and alpine tundra areas normally support a Lepidoptera fauna that, although relatively poor in species, is rich in numbers. After rains, deserts are often alive with butterflies and moths, but tropical forests are by far the richest in species. One of the strangest habitats occupied by a lepidopteran is the Neotropical three-toed sloth, on which the larvae of the sloth moth (*Bradypodicola hahneli*; Pyralidae) live, feeding on the algae that grow in the sloth's hair. The following list indicates some habitats of larval Lepidoptera.

1. **Plant associations, terrestrial**
   a. Feeding exposed on foliage, flowers, or plant lice
   b. Feeding on foliage in web or case
   c. Leaf and needle mining
   d. Living under bark
   e. Boring in stem, root, fruit, or seeds
   f. In soil, feeding on roots
   g. Boring in or feeding on fungi, mosses, liverworts, club mosses, lichens, and ferns
   h. Living in dried plant products, such as cereals, flour, dried fruit, refuse, and herbarium specimens

2. **Associations with social insects**
   a. In beehives

TABLE 2. Important species of Lepidoptera injurious to crops

Family and species	Damage by larvae
Sesiidae	
Peach-tree borer (*Synanthedon exitiosa*)	Feeds under bark; the most important insect injurious to peach
Squash-vine borer (*Melittia satyriniformis*)	Infests various cucurbits; often very destructive
Gelechiidae	
Angoumois grain moth (*Sitotroga cerealella*)	Feeds on grain in storage and in the field; imported from Europe
Pink bollworm (*Pectinophora gossypiella*)	Feeds on cotton bolls, and causes failure of blossoms to open; most important insect injurious to cotton
Tineidae	
Case-making clothes moth (*Tinea pellionella*); webbing clothes moth (*Tineola bisselliella*)	Clothes moth larvae infest woolen products, upholstery, furs, and other dried animal products
Carpet moth (*Trichophaga tapetzella*)	Damage similar to that by clothes moths but less common
Tortricidae	
Spruce budworm (*Choristoneura fumiferana*)	Defoliator of vast areas of coniferous forest
Ugly-nest tortricids, leaf rollers, etc. (*Cacoecia* spp., *Tortrix* spp.)	Feed on leaves and fruits of apples and other fruits, strawberry, shade trees, ornamental plants, etc.
Codling moth (*Cydia pomonella*)	Larva bores in apples and some other fruits; most important pest of apples
Oriental peach moth (*Cydia molesta*)	Feeds on fruits and twigs of peaches, plums, etc.
Pyralidae	
Bee moth (*Galleria mellonella*)	Destroys combs in neglected beehives
Indian meal moth (*Plodia interpunctella*)	Widespread and very important pest of dried fruit and animal products
Mediterranean flour moth (*Anagasta kuehniella*)	Infests flour, stored grain, cereals
Crambidae	
Oriental rice borer (*Chilo simplex*)	Very destructive to rice in Asia
Grape leaf folder (*Desmia funeralis*)	Sometimes an important defoliator of grapevines
Geometridae	
Spring cankerworm (*Paleacrita vernata*), fall cankerworm (*Alsophila pometaria*)	Defoliate fruit and shade trees in outbreak years; spring cankerworm overwinters as a pupa, fall cankerworm overwinters as an egg
Sphingidae	
Tomato hornworm (*Manduca quinquemaculatus*); tobacco hornworm (*M. sextus*)	Large larvae are most conspicuous pests of tomato and tobacco; also feed on other solanaceous plants
Lasiocampidae	
Tent caterpillars (*Malacosoma* sp.)	Serious defoliators of forest, shade, and orchard trees
Arctiidae	
Fall webworm (*Hyphantria cunea*)	Feeds on wide variety of forest, shade, and fruit trees
Lymantriidae	
European gypsy moth (*Lymantria dispar*)	Important defoliator of forest, shade, and fruit trees in New England
Brown-tail moth (*Euproctis chrysorrhoea*)	Similar to gypsy moth
Noctuidae	
Cutworms (*Euxoa* sp., *Peridroma* sp., etc.)	Numerous species damage many crop plants
Army worm (*Pseudaletia unipunctata*)	Varies tremendously in abundance; in epidemic years it does great damage to corn and other grasses
Corn earworm, cotton bollworm, tomato fruitworm, tobacco budworm (*Helicoverpa armigera*)	Exceedingly important cosmopolitan pest attacking many cultivated plants
Pieridae	
Cabbage butterflies (*Pieris* spp., especially *P. rapae* in United States and Europe)	Pests of various cultivated crucifers
Alfalfa butterfly (*Colias eurytheme*)	Important pest of alfalfa in southwestern United States

b. In ant nests
c. Possibly in termite nests

3. **Associations with other animals**
a. In stored animal products, such as woolens and feathers
b. In hair of three-toed sloths, feeding on algae
c. In plant lice, as parasitoids

4. **Aquatic associations**
a. Feeding on, boring in, or mining in aquatic plants
b. On rocks, feeding on microorganisms

In temperate zones, different species hibernate at every stage of the life cycle. Individual, seasonal, and geographic variations are common and striking in the Lepidoptera. Color and size are characters most frequently affected, but wing shape, venation, genitalia, and other structures can also be involved. The number of broods per season often differs with geographic location.

Behavior and physiology. Migration is the most spectacular behavior occurring in the order. It is most frequent in the butterflies, such as *Phoebis*, *Danaus*, and *Libytheana*, but is also known in moths such as *Chrysiridia* and the bogong moth of Australia (*Agrotis infusa*). Huge migratory swarms of Lepidoptera are frequently reported in many parts of the world; but this phenomenon, as well as the related communal roosting of adults, daily use of flyways, hill topping, and so forth, is poorly understood. Aggregations of butterflies at mud puddles are a frequent sight, and both moths and butterflies have been observed pumping—sipping water steadily from a puddle and ejecting it as a stream of drops or fine spray from the anus.

Evolution and genetics. There is no extinct family in this order, but it must be acknowledged that the known fossil record is quite poor. The family Eosetidae, based on *Eoses triassica* from the Triassic of Australia, was originally placed in the Lepidoptera, but its position is very doubtful. The earliest fossil lepidopteran is a homoneurous micromoth from the Lower Jurassic (*Archeolepis mane*), and so it seems possible that the order did first differentiate in the Triassic. *Archeolepis* was originally placed in its own (extinct) family, but this is no longer accepted, mainly because not enough characters can be appreciated to reliably locate it in, or exclude it from, one of the living families.

The phenomena of mimicry and protective resemblance are widespread in the Lepidoptera. It has been demonstrated that eyespot patterns in certain species elicit escape responses in passerine birds and are thus an effective protective device.

Lepidoptera have advantages for many kinds of genetic investigation, especially for the study of populations. The appearance of dark forms of various moths in heavily industrialized areas that have been polluted and blackened by soot is a widespread phenomenon. While this industrial melanism, most extensively studied in the geometrid *Biston betularia*, is now known to involve complex and shifting selection pressures, it remains one of the best-known examples of evolution in action. *See* POPULATION GENETICS.

Paul R. Ehrlich; R. I. Vane-Wright

Bibliography. C. L. Boggs, W. B. Watt, and P. R. Ehrlich (eds.), *Butterflies: Ecology and Evolution Taking Flight*, 2003; M. M. Douglas, *The Lives of Butterflies*, 1986; J. D. Holloway, G. Kibby, and D. Peggie, *The Families of Malesian Moths and Butterflies*, 2001; N. P. Kristensen (ed.), *Lepidoptera, Moths and Butterflies*, 2 vols., *Handbook of Zoology*, vol. 4 (parts 35, 36), 1999/2003; H. F. Nijhout, *The Development and Evolution of Butterfly Wing Patterns*, 1991; R. Preston-Mafham and K. Preston-Mafham, *Butterflies of the World*, 1988; R. Robinson, *Lepidoptera Genetics*, 1971; M. J. Scoble, *The Lepidoptera*, 1992; N. Wahlberg et al., Synergistic effects of combining morphological and molecular data in resolving the phylogeny of butterflies and skippers, *Proc. Roy. Soc. London B*, 272:1577–1586, 2005; M. Young, *The Natural History of Moths*, 1997.

Lepidosauria

A subclass of living and extinct diapsid reptiles. Lepidosauria and their immediate ancestors constitute the Lepidosauromorpha, one of the two major clades of the Diapsida. By definition, Lepidosauria includes the last common ancestor of the living squamates (lizards, snakes, and the limb-reduced burrowing amphisbaenians) and the New Zealand tuatara (*Sphenodon*) and all of its descendants. This is a narrower definition than that used in older literature, where Lepidosauria was used as a catchall group for all non-archosaurian diapsids. Lepidosaurs differ from archosaurs in that the lower temporal arcade (the inferior border of the lower temporal fenestra) is typically incomplete, there is never an antorbital fenestra in front of the orbit or a mandibular fenestra in the lower jaw, and the teeth are generally fused in position (acrodont or pleurodont) rather than implanted in sockets (thecodont). Other lepidosaurian characters include a specialized skin-shedding mechanism, paired male hemipenes (copulatory organs that characterize lizards and snakes), fracture planes in the caudal vertebrae that allow the tail to be shed if grabbed by a predator, and specialized knee, foot, and ankle joints that improve locomotion. As in mammals, the ends of lepidosaurian long bones develop separate centers of ossification (epiphyses) that fuse to the shaft at the end of skeletal growth. *See* ARCHOSAURIA; DIAPSIDA; REPTILIA.

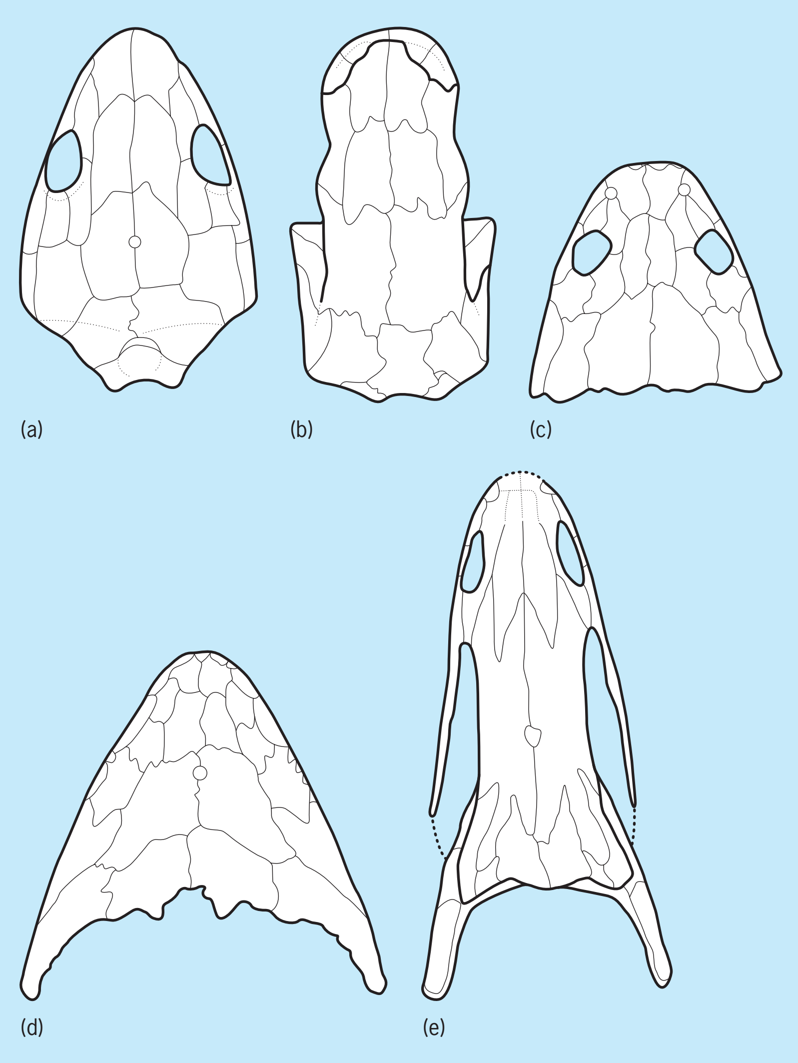
Characteristics. Living lepidosaurs can be divided into two subequal groups, Rhynchocephalia and Squamata. Abundant and globally diverse during the early Mesozoic, Rhynchocephalia are represented today only by the endangered tuatara (*Sphenodon*) of New Zealand. *Sphenodon* shows some primitive traits in comparison with squamates (such as rudimentary hemipenes, retention of belly ribs, and little or no cranial kinesis), but it is not a “living fossil” as once thought. The skull and dentition are specialized and the eardrum has been lost secondarily. Squamata form the largest and most successful group of living reptiles, with more than 6000 species and a global distribution. Many squamates are small insect eaters, but the group encompasses herbivores and major predators. They range in size from a few millimeters (such as some tiny geckos and chameleons) to several meters (Komodo dragons and anacondas), and in locomotor type they vary from quadrupedal to bipedal, from limbed to limbless, and from terrestrial runners to swimmers, specialized climbers, and gliders. Reproductive strategies range from egg laying to live birth, and at least eight groups have developed one or more all-female (parthenogenetic) species. *See* RHYNCHOCEPHALIA; SQUAMATA.

Fossil record. Most lepidosaurs are small with lightly built skeletons. Such skeletons are preserved and fossilized only under suitable conditions (fine sediments, quiet depositional environment), and their subsequent recovery depends on the use of an appropriate search strategy. Consequently, most lepidosaurian fossils have been recovered either in the search for early mammals (microvertebrate assemblages from caves, fissures, pond deposits) or early birds (fine-grained “lithographic” limestones), and there are major gaps in the record. The ancestors of squamates and rhynchocephalians probably diverged in the Early Triassic, but the first confirmed records are of rhynchocephalians from the Upper Triassic of Britain and Germany. These are already too derived to be an ancestral type, suggesting rhynchocephalians had a significant unrecorded Early to Middle Triassic history. The record of squamates is rather similar. The earliest known lizards are of Early to Middle Jurassic age (India, Central Asia, Britain), but they include representatives of several major lineages. This implies an earlier, Triassic beginning to the squamate radiation. Susan E. Evans

Bibliography. M. J. Benton, *Vertebrate Palaeontology*, 3d ed., Blackwell, Oxford. 2004; M. J. Benton (ed.), *The Phylogeny and Classification of the Tetrapods*, vol.1: *Amphibians, Reptiles and Birds*, Clarendon Press, Oxford. 1988; T. R. Halliday and K. Adler (eds.), *The New Encyclopedia of Reptiles and Amphibians*, Oxford University Press, 2002; G. R. Zug, L. J. Vitt, and J. P. Caldwell, *Herpetology*, Academic Press, San Diego, 2001.

Lepospondyli

An assemblage of diverse, extinct, small tetrapods found in rocks from the Mississippian to Permian periods (340–245 million years ago). In older Linnean classifications, Lepospondyli is a subclass of Amphibia, along with Labyrinthodontia (other Paleozoic tetrapods) and Lissamphibia (the modern amphibians). Recent analyses of interrelationships of all Paleozoic tetrapods have changed this view. Labyrinthodontia, for instance, has been found not to include all descendant lineages and has been abandoned as a technical name. However, research demonstrates that Lepospondyli includes all descendants of a single common ancestor (a clade), and continues to be used in both Linnean and phylogenetic classifications. Anatomical features uniting lepospondyls include a lack of labyrinthine infolding of tooth dentine, paired palatal fangs and replacement pits, absence of an otic notch in the back of the skull, a small forwardly oriented projection on the first vertebra (reduced in some), and single spool-shaped vertebrae (see **illustration**). However, these features are also characteristic of juvenile or small amphibians. Since all lepospondyls are small, a few specialists still question the group's integrity. *See* AMPHIBIA; LABYRINTHODONTIA; LISSAMPHIBIA.



Skulls of lepospondyls. (*a*) *Cardiocephalus*, a microsaur. (*b*) *Brachydectes*, a lysorophian. (*c*) *Scincosaurus*, a nectidean. (*d*) *Batrachiderpeton*, a nectridean. (*e*) *Coloraderpeton*, an aistopod. (*From J. S. Anderson, 2007*)

Taxonomy and characteristics. In Linnean taxonomy, Lepospondyli comprises five diverse orders: Microsauria, Nectridea, Lysorophia, Aistopoda, and Adelospondyli. All are also considered to be clades with the exception of Microsauria, which might include members more closely related to other lepospondyls than other microsaurs. Microsaurs are small, salamander- and lizardlike animals with diverse groupings (11 families) of variously elongate or consolidated body plans and a distinctive cranial articulation. Nectrideans are mostly aquatic, newtlike animals with unique tail vertebrae which support a broad fin for swimming. Adelospondyls, lysorophians, and aistopods are progressively more elongate animals with reduced limbs; aistopods lack limbs entirely, retain only a trace of the shoulder girdle, and have over 200 vertebrae. All are found in North America or Europe (Euramerica), with the exception of one late-occurring nectridean found in Morocco. *See* AISTOPODA.

Paleontology. Lepospondyls are receiving increased attention from paleontologists in recent years for two reasons. First, multiple studies indicate that lepospondyls are the last Paleozoic tetrapod group to share a common ancestor with the amniote lineage, so they are important for understanding the early evolution of amniotes. Second, some suggest that Lepospondyls gave rise to some or all of the modern amphibians. One hypothesis is that lissamphibians share a unique common ancestor closely related to lysorophians instead of another group of Paleozoic tetrapods, the temnospondyls. Another hypothesis is that caecilians are descended from microsaurs, but salamanders and frogs are descended from temnospondyls. A third hypothesis places salamanders and caecilians within microsaurs but leaves the origin of frogs within temnospondyls. These lepospondyl hypotheses for modern amphibian origins are minority views but are driving much new research into the anatomy of amphibians, including, for the first time, comparisons of development patterns between fossil and modern amphibians. *See* TEMNOSPONDYLI. Jason S. Anderson

Bibliography. J. S. Anderson, Incorporating ontogeny into the matrix: A phylogenetic evaluation of developmental evidence for the origin of modern amphibians, in J. S. Anderson and H.-D. Sues (ed.), *Major Transitions in Vertebrate Evolution*, Indiana University Press, Bloomington, in press, 2006; M. J. Benton, *Vertebrate Palaeontology*, 3d ed., Blackwell Press, London and New York, 2004; R. L. Carroll, *Vertebrate Paleontology and Evolution*, W. H. Freeman, New York, 1988.

Leprosy

A chronic infectious disease caused by *Mycobacterium leprae* that primarily affects the skin and peripheral nerves and, to a lesser extent, involves the eyes and mucous membranes. Leprosy, or Hansen's disease, has been known for more than 2000 years. It afflicts at least 3 million people worldwide and is most common in developing countries. There are about 200 new cases yearly in the United States, 85% of them among aliens. The pathogen, *M. leprae*, has never been cultured in artificial media, but it can be grown in the mouse (in the footpad) and the nine-banded armadillo, among other animals. About 2–3% of armadillos in the wild in the southern United States harbor the infection. Its epidemiology is not fully understood, but transmission probably takes place by the respiratory route. The bacillus is very slow-growing, and the incubation period is usually 3–5 years. Less than 5% of any population is susceptible, and these individuals have a deficient cell-mediated immune response specifically to *M. leprae*, which may be genetic in origin. Epidemics have occurred, but are rare. Historically, considerable stigma has been attached to the disease and remains a real, though diminishing, problem. *See* MYCOBACTERIAL DISEASES.

A skin rash and loss of feeling due to nerve damage by *M. leprae* are the hallmarks of leprosy. Usually the nerve damage is mild, but when severe the inability to feel, particularly in the hands and feet, predisposes the individuals to frequent injuries. Nerve involvement may also lead to a loss of muscle function that produces clawing of the fingers and toes as well as other neuromuscular dysfunctions. Manifestations of the disease depend upon the degree of the immune defect. Initially, most patients develop one or several depigmented areas of skin that may have decreased sensation, a stage referred to as indeterminate disease. The condition may self-heal, but if it is diagnosed it is always treated. If treatment or self-healing does not halt the disease, it may progress to one of three advanced types. The mildest of these, tuberculoid disease, is usually manifested as a single large depigmented, scaly, numb area. The most severe type, lepromatous leprosy, usually involves most of the skin to varying degrees, with variously sized nodules or other changes, but some cases have no distinct rash. Between the two extremes immunologically is borderline disease, with skin changes of both types (see **illus.**). The World Health Organization's simplified classification labels indeterminate and tuberculoid patients paucibacillary (few bacilli) and borderline and lepromatous patients multibacillary. Diagnosis is confirmed by a finding of bacteria in material scraped from tiny skin slits or a skin biopsy that shows the presence of bacilli and nerve involvement.

Borderline leprosy showing skin changes ranging from large tuberculoid lesions to lepromatous nodules.

Paucibacillary disease is treated with dapsone plus rifampin for 6–12 months. Clofazimine is added for multibacillary individuals and the treatment continued for 2 or more years. Treatment may be complicated by reactive episodes marked by fever, skin changes, and nerve inflammation. These side effects can usually be managed with drugs such as corticosteroids and thalidomide. Isolation is unnecessary, since patients become noninfectious within days of starting treatment. All can, in effect, be cured. *See* INFECTIOUS DISEASE.

Robert R. Jacobson

Bibliography. R. B. Conn (ed.), *Current Diagnosis 9th ed.*, 1999; H. D. Humes (ed.), *Textbook of Internal Medicine*, 4th ed., 2000; R. E. Rakel (ed.), *Conn's Current Therapy*, 1999.

Leptin

A protein comprising 167 amino acids that is important in the regulation of body weight, metabolism, and reproduction. Leptin was discovered in 1994, when it was identified as the missing protein in mice with a spontaneous single-gene defect that caused obesity. These *ob/ob* mice were very obese, had many of the metabolic abnormalities associated with obesity, and were infertile. However, when leptin levels were examined in other animal models of obesity and in obese humans, the levels were found to be elevated and not low or absent (as in the *ob/ob* mice). Thus, general obesity (not due to leptin deficiency) has been termed a leptin-resistant state. Since these initial findings, the biology of leptin has proven to be more complex than originally thought.

Secretion. Leptin is secreted primarily from adipocytes (fat cells), although other cells in other locations in the body, including the placenta, stomach, and skeletal muscle, also make leptin. Leptin is secreted in a pulsatile fashion, and the average levels vary over the day (with the lowest levels in midmorning and the highest levels at night). The amount of leptin secreted by adipocytes is in proportion to the size of the cells. In animals, leptin levels increase soon after a meal and decrease with fasting. In humans, leptin levels do not change significantly with normal meals or with reasonable levels of exercise. However, with massive overfeeding, leptin levels will increase by up to 40%, and with fasts longer than 24 hours, leptin levels will decrease as much as 50%. Thus, in humans, leptin appears to play a role in the more chronic regulation of body fat rather than the meal-to-meal changes in energy intake and energy expenditure.

Insulin requirement. Insulin is required for leptin secretion. In animals and in cell culture, insulin increases leptin expression. Likewise, leptin levels decrease when animals are made insulin-deficient. There is also a correlation between insulin levels and leptin levels in humans, but it is difficult to separate the confounding effect of increasing adipose tissue. When humans are deficient in insulin, leptin levels are low, but body fat is also decreased. However, when humans are given excess insulin for several hours (while keeping the blood glucose normal), most studies find no changes in leptin levels. Other hormones that increase leptin secretion include glucocorticoids (such as prednisone) and cytokines (such as tumor necrosis factor alpha and interleukin-1). Hormones that decrease leptin secretion include catecholamines (such as isoproterenol and epinephrine), testosterone, and perhaps thyroid hormone.

Gender difference. The leptin level in women is almost three times higher than the leptin level in men. One reason is that women have a higher percent of body fat than men. Additionally, women have more subcutaneous fat (they tend to be "pear-shaped") than men (they tend to have more abdominal fat and are "apple-shaped"). Subcutaneous fat produces more leptin than abdominal fat, thus contributing to this gender difference. Finally, testosterone, which is present in much greater concentrations in men, decreases leptin and further augments the gender difference.

Receptors. The primary role of leptin appears to be in decreasing food intake by way of its action in the brain. Leptin's actions are exerted by binding to specific receptors. There are at least six different splice variants (isoforms) of the receptor for leptin. All of these receptors contain a leptin-binding area, and all but one receptor span the cell membrane. However, only one receptor (the long-receptor isoform, or Ob-Rb) has complete signaling function and is found predominantly in the hypothalamus (in areas known to control feeding behavior and hunger). The short-receptor isoforms are located mainly in peripheral tissues (such as the lung, kidney, liver, and gonads) and the choroid plexus, but the current role of these short forms of the leptin receptor is still under investigation. They likely play a role in leptin transport in the blood and across the blood-brain barrier. At least part of the resistance to leptin associated with obesity might be due to decreased transport across the blood-brain barrier.

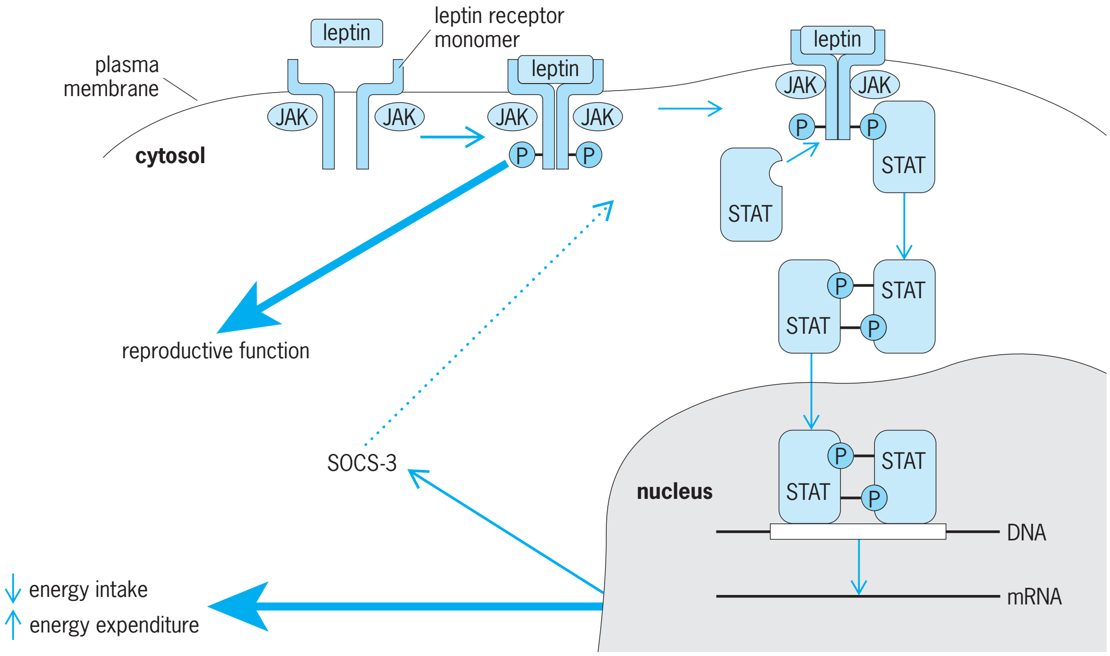
JAK-STAT signaling pathway. When leptin binds to the long form of the receptor, the receptor dimerizes (two receptors come together in combination with leptin) and several cascades are initiated by way of activation of the JAK (janus kinase)–STAT (signal transducers and activators of transcription) pathway (see **illus.**). This activation by JAK results in phosphorylation of tyrosine amino acids on the leptin receptor, which causes changes in gene expression (via STAT pathways that ultimately decrease appetite and increase energy expenditure), downstream phosphorylation and dephosphorylation of enzymes, alterations in ion channels, and other signal modulation (in part via SOCS-3, a suppressor of cytokine signaling). SOCS-3 can actually feed back to the leptin receptor and inhibit leptin signaling, and this may also play a role in the leptin resistance seen in obesity.

It has recently been shown that the effects of leptin on appetite occur through the STAT pathway, but the effects on maintaining reproductive hormones require only the JAK and not the STAT pathway. Additionally, activation of the leptin receptor can also activate pathways analogous to those of the insulin receptor, including the PI-3 kinase pathway (phosphatidyl inositol kinase pathway, which stimulates glucose uptake and other metabolic effects of insulin) and the MAP kinase pathway (mitogen-activated kinase pathway, which stimulates cell growth).

Regulation of body weight. Leptin is believed to regulate body weight through its effects on appetite and energy expenditure. When leptin is given to animals, there is a marked decrease in appetite and an increase in energy expenditure.

Neural pathways of appetite regulation. The decrease in appetite associated with leptin administration is due to the stimulation of neurons that inhibit feeding and the inhibition of neurons that promote feeding. The long form of the leptin receptor is found on two distinct types of neurons, the POMC/CART (proopiomelanocortin and cocaine- and amphetamine-regulated transcript) neurons and the NPY/AGRP (neuropeptide-Y and agouti-related peptide) neurons. The POMC/CART neurons decrease food intake when stimulated by neurotransmitters such as leptin and α-MSH (melanocyte-stimulating hormone, a product of POMC that is made in the hypothalamus and binds to the melanocortin receptors on these neurons). Leptin also binds directly to the NPY/AGRP neurons and decreases food intake by inhibiting the neurons' ability to be stimulated. However, NPY and AGRP stimulate feeding. NPY stimulates food intake by binding directly to receptors on the NPY/AGRP neurons, whereas AGRP stimulates food intake indirectly by antagonizing or counteracting the effect of α-MSH on the melanocortin receptors.

Effects on energy expenditure. The increase in energy expenditure is due to stimulation of the sympathetic



Leptin signal cascade. Leptin binds to the long form of the receptor, resulting in receptor dimerization and initiating several signaling cascades. The primary cascade signals by way of activation of the JAK-STAT pathway. The activation by JAK results in phosphorylation of tyrosine amino acids (shown as P in circle) on the leptin receptor. This in turn results in changes in gene expression via STAT pathways, which ultimately decrease appetite and increase energy expenditure. The maintenance of reproductive function requires only the JAK and not the STAT pathway. Signal modulation occurs in part via SOCS-3, a suppressor of cytokine signaling. SOCS-3 can feed back to the leptin receptor and inhibit leptin signaling. (*Adapted from D. L. Nelson and M. M. Cox*, *Lehninger Principles of Biochemistry*, *3d ed.*, *Worth Publishers*, *New York*, *2000*)

nervous system, which causes increases in oxygen consumption and metabolic rate. In addition, leptin has been shown to induce adipose cell death (apoptosis). As animals become more obese, the response to leptin given peripherally (into the abdomen) decreases, and they become leptin-resistant. Although the ability to respond to central leptin administration (leptin given directly into the brain) remains intact, with continued obesity there is also a decrease in the response to central leptin administration. This adipose cell death, along with an alteration of how the body handles nutrients following leptin administration, might be why there is a preservation of muscle with weight loss (in contrast to weight loss from caloric restriction alone, which is accompanied by a loss in muscle).

Regulation of reproductive hormones. When leptin-deficient mice were first discovered, they were found to be infertile as well as obese. When leptin was given back to these animals, fertility was restored. In addition, when leptin is given to animals prior to puberty, they undergo puberty earlier than control animals. Thus leptin appears to be part of the system that tells the brain there are adequate energy stores (adipose tissue) for reproduction and allows puberty to commence. Leptin has been shown to increase the secretion of the hormones that regulate reproductive function, including gonadotropin releasing hormone, luteinizing hormone, and follicle stimulating hormone.

Other potential roles. Although most of the attention of leptin has centered on its role in decreasing food intake and decreasing weight, it is not certain that this is its primary role. Until very recently, there has been little need for a hormone that inhibits an organism from gaining as much weight as possible, for more efficient weight gain appears to protect against the next famine. Therefore, leptin likely has other roles as a hormone.

Fat accumulation. One role that has received much attention lately is that of a hormone that prevents fat accumulation in nonadipose tissues, including muscle, the liver, and the pancreas. Excess fat accumulation has been shown to be detrimental to these organs and may play a significant role in the development of diabetes. Leptin may traffic fat in such a way that it does not accumulate in these tissues. In addition, leptin increases utilization of fat by producing heat and not energy. Thus leptin might enhance more appropriate and rapid disposal of fat that enters these tissues.

Fasting. Another potential role of leptin is in mediating the endocrine response to fasting. When a human or mouse is fasted, there is a decrease in levels of leptin, reproductive hormones, thyroid hormone, and insulin. However, there is an increase in levels of glucocorticoids, catecholamines, and glucagon. This results in a conservation of energy, an increase in the production of glucose for the brain, and an increase in the release of fatty acids from the adipose tissue for muscle. When mice were fasted but given leptin, there was a blunting in several of these responses, including the changes in levels of glucocorticoids, reproductive hormones, and thyroid hormone. Thus rather than acting as a signal of excess adipose tissue, leptin's primary role might be in signaling a decrease in energy, and initiating the stress response to this decrease.

Blood cell and bone formation. There is currently much debate on the extent of direct leptin action outside the brain. The gastrointestinal tract does have the long form of the leptin receptor, and signals from this receptor might play a role in nutrient absorption or appetite regulation. In addition, there is evidence that leptin has a direct role in the formation of bone red blood cells, white blood cells, and blood vessels.

Treatment of obesity. There are only five known people (from two families) who are deficient in leptin (like the *ob/ob* mice). They are very obese and seem to have reproductive abnormalities, but they do not appear to have other abnormalities such as increased glucose or insulin. Treatment with leptin has had a dramatic impact on their weight and has allowed puberty to occur. There are also three people (from one family) who do not make normal leptin receptors. These individuals have very high leptin levels, are very obese, do not have normal reproductive function, and have abnormal levels of thyroid hormone and growth hormone. Similar to the patients with leptin deficiency (and unlike the mouse models), they do not have high levels of blood glucose.

Leptin is also under investigation as a treatment for obesity in people with elevated leptin levels. Initial reports suggest that leptin administration results in weight loss in these subjects (with preservation of muscle), with the degree of weight loss similar to other pharmacologic treatments for obesity (about 8%). However, since leptin is a protein, it cannot be taken by mouth and must be injected, which sometimes results in pain, redness, inflammation, and bruising at the injection site. This mode of administration and the injection site reactions limit its use as a treatment. Whether administration of leptin or of a leptin-like compound will ever become a treatment for obesity remains to be seen. However, the discovery of leptin has opened a new area in the understanding of obesity and body weight regulation. *See* DIABETES; ENDOCRINE SYSTEM (VERTEBRATE); HUNGER; INSULIN; LIPID METABOLISM; METABOLIC DISORDERS; OBESITY.

William T. Donahoo; Robert H. Eckel

Bibliography. R. S. Ahima and J. S. Flier, Leptin, *Annu. Rev. Physiol.*, 62:413–437, 2000; C. A. Baile, M. A. Della-Fera, and R. J. Martin, Regulation of metabolism and body fat mass by leptin, *Annu. Rev. Nutrit.*, 20:105–127, 2000; R. B. Harris, Leptin—much more than a satiety signal, *Annu. Rev. Nutr.*, 20:45–75, 2000; B. Jeanrenaud and F. Rohner-Jeanrenaud, Effects of neuropeptides and leptin on nutrient partitioning: Dysregulations in obesity, *Annu. Rev. Med.*, 52:339–351, 2001; D. L. Nelson and M. M. Cox, *Lehninger Principles of Biochemistry*, 3d ed., Worth Publishers, New York, 2000; R. H. Unger, Leptin physiology: A second look, *Regulatory Peptides*, 92(1–3):87–95, 2000.

Lepton

An elementary particle having no internal constituents. There are three known charged leptons: the electron (e^-), the muon (μ^-), and the tau (τ^-). In addition, the corresponding antiparticles e^+, μ^+, τ^+ are known. Each experiences electromagnetic, weak, and gravitational forces, but not the strong (nuclear) force. Associated with each charged lepton is a corresponding neutral lepton, called a neutrino (ν_e, ν_μ, ν_τ) or antineutrino: ($\bar{\nu}_e$, $\bar{\nu}_\mu$, $\bar{\nu}_\tau$); these have only weak and gravitational interactions. A charged lepton and its associated neutrino form a lepton generation; thus there are three known lepton generations (see **table**). *See* ELECTRON; NEUTRINO.

Leptons are very small, less than 10^{-18} m in radius. This is less than 10^{-3} of the radius of a nucleus and less than 10^{-8} of the radius of an atom. Indeed, all existing measurements are consistent with the assumption that leptons are point particles.

The lepton family of particles has very different properties than the quark family of particles. Quarks interact through the strong force as well as through the electromagnetic, weak, and gravitational forces. By means of the strong force, quark-antiquark pairs bind together to form mesons such as the π meson, and three quarks bind together to form baryons such as the proton. Whereas, as far as is known, quarks are always confined inside hadrons (mesons and baryons) and cannot be studied as isolated particles, leptons act as individual particles and can be studied as isolated particles. *See* BARYON; FUNDAMENTAL INTERACTIONS; HADRON; MESON; QUARKS.

More charged leptons or neutrinos may exist. Searches have been conducted up to 45 GeV in mass for additional charged leptons or neutrinos, but none have been found. Existing accelerators are not powerful enough to allow searches above 45 GeV. The reason there are only three lepton generations below 45 GeV in mass (and that there are only three generations of quarks) is unknown.

Decay. The electron is a stable particle: it never decays. The muon and tau are unstable particles, and their average lifetimes have been measured (see table). There is only one known mode of muon decay [reaction (1)].

$$\mu^- \rightarrow \nu_\mu + e^- + \bar{\nu}_e \qquad (1)$$

The tau is heavier than the muon and thus decays more rapidly and in many modes, as in reactions (2).

$$\begin{aligned} \tau^- &\rightarrow \nu_\tau + \mu^- + \bar{\nu}_\mu \\ \tau^- &\rightarrow \nu_\tau + e^- + \bar{\nu}_e \\ \tau^- &\rightarrow \nu_\tau + \pi^- \end{aligned} \qquad (2)$$

Lepton conservation and lepton mass. The association between charged leptons and their neutrinos has been described by an empirical law called lepton conservation. For example, consider the electron (e^-). There are only two ways in which an e^- can be destroyed: it can be annihilated by combining it with an e^+ (positron); or it can be changed into a ν_e. But an e^- cannot be changed directly into any other

Details of the three lepton generations

	Generation 1	Generation 2	Generation 3
Charged lepton name	Electron	Muon	Tau
Charged lepton symbol	$e^{\pm}$	$\mu^{\pm}$	$\tau^{\pm}$
Dates of discovery	1890s	Late 1930s	1974–1975
Charged lepton mass, MeV	0.51	105.7	~1777
Charged lepton lifetime, s	Stable	2.2×10^{-6}	3.0×10^{-13}
Associated neutrino symbols	ν_e, $\bar{\nu}_e$	ν_μ, $\bar{\nu}_\mu$	ν_τ, $\bar{\nu}_\tau$

lepton or hadron, such as an e^+, μ^-, μ^+, τ^-, τ^+, or π. Furthermore an e^- cannot be changed into a muon-associated neutrino (ν_μ, $\bar{\nu}_\mu$) or into a tau-associated neutrino (ν_τ, $\bar{\nu}_\tau$). Thus it seems that there is a unique property of the electron (e^-) which is preserved or conserved in all reactions; the only other particle carrying this unique property is the electron-associated neutrino (ν_e).

Similar lepton conservation laws appear to hold for the μ and τ. The only other particle carrying the unique property of the μ^- is the muon-associated neutrino (ν_μ). The only other particle carrying the unique property of the τ^- is the tau-associated neutrino (ν_1).

Until 1998 all experimental evidence was consistent with the assumption that lepton conservation is universally valid. However, it is now known that the law breaks down in the phenomenon of neutrino oscillations, which occurs because (1) the neutrinos ν_e, ν_μ, ν_τ do not have definite masses; (2) a quantum-mechanical mixing of neutrino states can occur. It is now known from the results of neutrino oscillation experiments that the quantum-mechanical state of, for example, ν_μ must be expressed as a specific linear combination of three neutrino mass states, at least two of which are associated with definite and distinct nonzero mass (and the same general statement is true of ν_e and ν_τ). Consider a ν_μ that is created at time $t = 0$ in muon decay [reaction (1)]. As this neutrino propagates through space, the states of definite and distinct mass of which it is composed evolve at different rates and thus get out of phase. The result is that after a finite time has elapsed, the neutrino is no longer a pure ν_μ but is in a superposition of ν_e, ν_μ, ν_τ states. Hence there is a finite probability that in a collision with another particle this neutrino could be converted to an electron or (if enough energy is available) even to a tau lepton. This violates lepton conservation.

The masses of the charged leptons (see table) are given in the energy equivalent unit megaelectronvolts (MeV). One MeV is the energy gained by an electron that is accelerated through a voltage of 10^6 V, and is equal to 1.602×10^{-13} joule. The masses of the electron, muon, and neutrinos are smaller than the masses of any of the hadrons. However, the discovery of the tau, whose mass is larger than that of many hadrons, destroyed the concept that leptons had to be very small mass particles.

The mass values m_1, m_2, m_3 of the neutrino states of definite mass are not yet known. At present, results of neutrino oscillation experiments reveal only that Eqs. (3) apply, where c is the velocity of light. These

$$\begin{aligned} m_2^2 - m_1^2 &= 8.1 \cdot 10^{-5} (\mathrm{eV}/c^2) \\ m_3^2 - m_1^2 &= 2.2 \cdot 10^{-5} (\mathrm{eV}/c^2) \end{aligned} \tag{3}$$

results imply that at least two of the three states are associated with nonzero (but certainly very small) mass.

Production. Leptons can be produced in various ways. One way is through the decay of a hadron. For example, a muon and its associated antineutrino are produced when a pion decays, as in reaction (4).

$$\pi^- \rightarrow \mu^- + \bar{\nu}_\mu \tag{4}$$

Another way to produce leptons is through the annihilation of an electron with a positron; for example, muons can be produced by reaction (5), and taus can be produced by reaction (6).

$$e^+ + e^- \rightarrow \mu^+ + \mu^- \tag{5}$$

$$e^+ + e^- \rightarrow \tau^+ + \tau^- \tag{6}$$

It was through reaction (6) that the tau was discovered. *See* ELEMENTARY PARTICLE.

Eugene D. Commins; Martin L. Perl

Bibliography. E. Commins and P. Bucksbaum, *Weak Interactions of Leptons and Quarks*, Cambridge University Press, 1983; F. Halzen and A. D. Martin, *Quarks and Leptons: An Introductory Course in Modern Particle Physics*, Wiley, New York, 1984; B. Okun, *Particle Physics: The Quest for the Substance of Substance*, Harwood, London, 1985.

Leptospirosis

A bacterial disease caused by a *Leptospira* species that is now recognized as one of the new, emerging infectious diseases. The Centers for Disease Control and Prevention (CDC) has referred to "new, reemerging, or drug-resistant infections whose incidence in humans has increased within the past two decades or whose incidence threatens to increase in the near future." Leptospirosis, presumed to be the most widespread zoonosis (a disease of animals that may be transmitted to humans) in the world, is characterized by a broad spectrum of clinical manifestations, varying from a subclinical infection (showing no symptoms) to a fatal disease.

Etiologic agents. Leptospires are spirochetes (flexible, spiral-shaped bacteria with flagella) that are coiled, thin, and highly motile; have hooked ends;

and contain two flagella, which enable the bacteria to move in a watery environment and burrow into and through tissue. Traditionally, the genus *Leptospira* contains two species: the pathogenic *L. interrogans* and the free-living *L. biflexa*. Leptospires are phylogenetically related to other spirochetes that have a genome approximately 5000 kilobases in size. Recently, a number of leptospiral genes have been cloned and analyzed.

Epidemiology. Leptospirosis is primarily an animal disease with a worldwide distribution that affects at least 160 mammalian species. The incidence is significantly higher in warmer climates than in temperate regions, mainly due to longer survival of leptospires under warm humid conditions. Animals can be divided into maintenance hosts and accidental hosts.

Maintenance hosts are species in which infection is endemic and is usually transferred from animal to animal by direct contact. Infection is usually acquired at an early age, and the prevalence of chronic excretion of the leptospires in the urine increases with the age of the animal. The disease is maintained in nature by chronic infection of the renal tubules of maintenance hosts. Rats, mice, opossums, skunks, raccoons, and foxes are the most important environmental maintenance hosts, with farm animals and dogs being domestic maintenance hosts. All of these animals shed the bacteria in their urine.

Humans are not commonly infected with leptospires and are considered an accidental (incidental) host. Approximately 100 human cases of leptospirosis are reported to the CDC each year, but this probably misrepresents the total number of cases.

Transmission. Transmission of leptospires may follow direct contact with urine, blood, or tissue from an infected animal or exposure to a contaminated environment. Human-to-human transmission is rare. Since leptospires are excreted in the urine, they can survive in water for many months. Leptospires can enter the host through abrasions in the skin or through mucous membranes such as the conjunctiva. Drinking contaminated water may introduce the bacteria through the mouth, throat, or esophagus. The incubation period is usually 1–2 weeks.

Certain occupational groups are at high risk for acquiring leptospires. These include veterinarians, sewage workers, agricultural workers, livestock farmers, slaughterhouse workers, meat inspectors, rodent-control workers, and fish workers. They acquire the bacteria by direct exposure or contact with contaminated soil or water. In western countries, there is a significant risk of acquiring leptospirosis from recreational exposure (canoeing, windsurfing, swimming, water skiing, white-water rafting, freshwater fishing) and domestic animal contact.

Pathogenesis. The mechanisms by which leptospires cause disease are not well understood. It is known that after the leptospires enter the host, they spread to all organs of the body. The bacteria reproduce in the blood and tissues. Although any organ can be infected, the kidneys and liver are most commonly involved. In these organs, the leptospires attach to and destroy the capillary walls. In the liver this leads to jaundice, and in the kidneys it causes inflammation (nephritis) and tubular cell death (necrosis), resulting in renal failure. After antibodies are formed, the leptospires are eliminated from all organ sites except the eyes and kidneys, where they may persist for weeks or months, for unknown reasons.

Clinical manifestations. The clinical manifestations of leptospirosis can be divided into two forms: anicteric (not jaundiced) and icteric (jaundiced).

The majority of infections caused by the leptospires are anicteric (either subclinical or of very mild severity), and patients usually do not seek medical attention. In this mild form, leptospirosis may be similar to an influenza-like illness with headache, fever, chills, nausea, vomiting, and muscle pain. Fever is the most common finding during a physical examination. Most patients become asymptomatic within 1 week.

Between 5 and 15% of all patients with leptospirosis have the icteric (severe) form of the disease, which is characterized by jaundice, renal and liver dysfunction, hemorrhaging, and multiorgan involvement, and can result in death. This severe form is also referred to as Weil's syndrome (after Adolf Weil, a German physician). The mortality rate for is 5–15%. The jaundice occurring in leptospirosis is not associated with hepatocellular necrosis, and liver function returns to normal after recovery. Acute infections in pregnancy have also been reported to cause abortion and fetal death.

The clinical presentation of leptospirosis is biphasic, with the acute or septicemic (characterized by bacteria in the blood) phase lasting about a week. This is followed by the immune phase, characterized by antibody production and excretion of the leptospires in the urine. Most of the complications of leptospirosis are associated with localization of the leptospires within the tissues during the immune phase and thus occur during the second week of the infection.

Laboratory findings. Since the kidneys are invariably involved in leptospirosis, related findings include presence of white blood cells, red blood cells, and granular casts (proteinaceous products of the kidney in the urine); mild proteinuria (protein in the urine); and in severe leptospirosis, renal failure and azotemia (an increase in nitrogenous substances in the blood). The most common x-ray finding in the kidney is a patchy pattern that corresponds to the many hemorrhages due to the destruction of blood vessels.

Diagnosis. A definitive diagnosis of leptospirosis is based either on the isolation of the leptospires from the patient or on immunological tests. Leptospires can be isolated from the blood and/or cerebral spinal fluid during the first 10 days of the illness, and from the urine for several weeks beginning about the first week of the infection. They can be stained using carbol fuchsin, counterstained, and cultured in a simple medium enriched with vitamins.

Treatment. Treatment of leptospirosis differs depending on the severity and duration of symptoms at the time of presentation. Patients with only flu-like symptoms require only symptomatic treatment but should be cautioned to seek further medical help if they develop jaundice. The effect of antimicrobial therapy for the mild form of leptospirosis is controversial. However, such treatment is definitely indicated for the severe form of the disease, and should begin as soon as a severe diagnosis is made. In mild cases, oral treatment with tetracycline, doxycycline, ampicillin, or amoxicillin may be considered. For more severe cases of leptospirosis, intravenous administration of penicillin G, amoxicillin, ampicillin, or erythromycin is recommended. Most patients with leptospirosis recover. Mortality is highest in the elderly and those who have Weil's syndrome. Leptospirosis during pregnancy is associated with high fetal mortality. Long-term follow-up of patients with renal failure and hepatic dysfunction has documented good recovery of both renal and liver function.

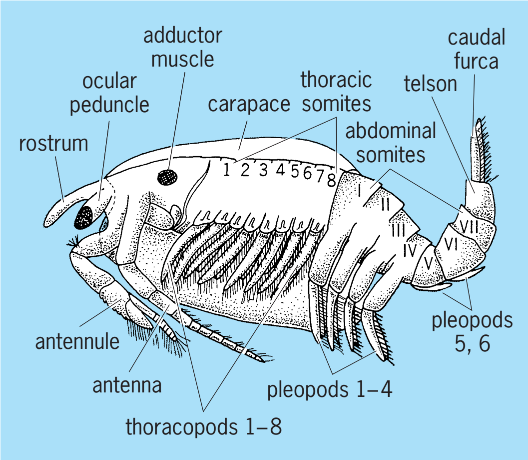
Prevention. Individuals who may be exposed to leptospires through either their occupation or involvement in recreational water activities should be informed about the risks. Some measures for controlling leptospirosis include avoiding exposure to urine and tissues from infected animals, vaccination of domestic animals, and rodent control. *See* BACTERIA; MEDICAL BACTERIOLOGY; SPIROCHETE; ZOONOSES.

John P. Harley

Bibliography. R. W. Farrar, Leptospirosis, *Crit. Rev. Clin. Lab. Sci.*, 21(1), 1995; R. W. H. Gillespie, Epidemiology of leptospirosis, *Amer. J. Pub. Health*, vol. 53, 1963; P. N. Levett, Leptospirosis, *Clin. Microbiol. Rev.*, 14(2), 2001; R. van Crevel, Leptospirosis in travelers, *Clin. Infect. Dis.*, 19:132, 2001.

Leptostraca

The only extant order of the crustacean subclass Phyllocarida. The Leptostraca is represented by one fossil and a small number of living genera. These malacostracans (see **illus.**) are unique in having the carapace laterally compressed to such an extent that it forms a bivalvelike shell held together by a strong adductor muscle. The carapace covers only the thorax, leaving exposed the head, with its uniquely movable rostrum, stalked eyes, paired antennules, and antennae. The thorax is composed of eight malacostracan somites, each with a pair of generally similar appendages (thoracopods). Unlike the typical malacostracan abdomen, the leptostracan abdomen consists of seven somites plus a telson with furcal rami. Only the first six abdominal somites carry paired appendages (pleopods). The retention of a seventh somite and caudal furca, in addition to a fully developed telson in the Leptostraca suggested an evolutionarily early offshoot from the malacostracan stem. Although leptostracans are the only extant malacostracans to have seven clearly delineated abdominal somites, evidence from embryological and neurological studies indicates that in some members of the Mysidacea a seventh somite, or the ganglia thereof, remains in developmental stages. *See* MALACOSTRACA.



***Paranebalia longipes*. (*After P. A. McLaughlin, Comparative Morphology of Recent Crustacea, W. H. Freeman, 1980*)**

Leptostracans use the thoracopods to produce a feeding current and, in females, to form a brood pouch. That secondary brooding function suggests that egg-bearing females generally do not feed, and evidence confirms this. Locomotion in leptostracans is accomplished by use of the first four pairs of pleopods. Most leptostracans are bottom dwellers, living on or slightly under the substrate, but one species is holopelagic, one inhabits hydrothermal vents, and still another is a marine cave dweller. *See* CRUSTACEA; PHYLLOCARIDA.

Patsy A. McLaughlin

Bibliography. T. E. Bowman, J. Yager, and T. M. Iliffe, *Speonebalia cannoni*, n. gen., n. sp., from the Caicos Islands, the first hypogean leptostracan (Nebaliacea: Nebaliidae), *Proc. Biol. Soc. Wash.*, 98:439–446, 1985; R. R. Hessler, *Dahella caldariensis*, new genus, new species: A leptostracan (Crustacea, Malacostraca) from deep-sea hydrothermal vents, *J. Crust. Biol.*, 4:655–664, 1984; P. A. McLaughlin, *Comparative Morphology of Recent Crustacea*, 1980; F. R. Schram, *Crustacea*, 1986.

Lespedeza

A warm-season legume with trifoliate leaves, small purple pea-shaped blossoms, and one seed per pod. There are 15 American and more than 100 Asiatic species; two annual species, *Lespedeza striata* and *L. stipulacea*, and a perennial, *L. cuneata*, from Asia are grown as field crops in the United States. The American species are small shrubby perennials (see **illus.**) found in open woods and on idle land, rarely in dense stands; they are harmless weeds. Common lespedeza, once known as Japanese clover, is a small

Lespedeza. ***(Soil Conservation Service)***

variety of *L. striata*. This variety, unintentionally introduced in the 1840s and used widely until the late 1920s, has been replaced by a larger variety, Kobe. Korean lespedeza (*L. stipulacea*) has been widely grown since the mid-1920s and is preferred in the northern part of the lespedeza belt, from Missouri eastward, while Kobe is preferred in the lower part of the belt (south to the Gulf of Mexico and across northern Florida). *See* LEGUME.

The annual lespedezas can grow on poorer wetter land than most other forage crops. They give a moderate yield of good quality. The plants reseed well and the young seedlings at the two-leaf stage are very tolerant of cold. Most crop damage is from parasitic nematodes and severe drought, especially in the spring.

The perennial lespedeza (*L. cuneata*), commonly known as sericea, is adapted to the same growing conditions as the annuals but persists better on steep slopes and for this reason has been used widely in soil conservation. The young growth from well-established sericea is a moderate-quality forage, but the older growth is poor. The perennial is difficult to establish because the seed does not germinate readily unless scarified or exposed to cold and moisture for several months.

Several large species of lespedeza, such as the widely used *L. bicolor*, are used in plantings for wild life to provide seed and cover. *See* LEGUME FORAGES; NEMATA (NEMATODA). Paul Tabor

Lethal dose 50

One special form of the effective dose 50, or median effective dose. The lethal dose 50, or median lethal dose, is also written as LD_{50}. It is used when the response that is being investigated is the death of the experimental animal. The median lethal dose is therefore the dose which is fatal to 50% of the test animals.

The LD_{50} for strychnine given subcutaneously to rabbits is 0.5 mg/kg of body weight. This description shows that in specifying the LD_{50} attention has been given to the species of animal, the route of injection, and the weight of the animals used. For other toxic substances a different set of specifications might be important.

The use of the concept of LD_{50} is not restricted to toxicology, because there are many therapeutic substances, such as digitalis, which can be assayed more their lethal effect than by their therapeutic effect. The same active principle is involved in both effects. *See* BIOASSAY; EFFECTIVE DOSE 50; TOXICOLOGY. Colin White

Lethal gene

A gene which brings about the death of the organism carrying it. Lethal genes constitute the most common class of gene changes (that is, mutations) and are reflections of the fact that the fundamental function of genes is the control of processes essential to the growth and development of organisms. In higher diploid forms lethals are usually recessive and expressed only in homozygotes. Dominant lethals, expressed in heterozygotes, are rapidly eliminated and thus rarely detected. Recessive zygotic lethals are retained with considerable frequency in natural populations of cross-fertilizing organisms, while gametic lethals (those affecting normal functioning of eggs and sperm among animals, or the pollen and ovules of plants) are subject to stringent selection and are accordingly rare.

Evaluation and analysis. Some lethal mutations have proved to be losses of small or large sections of chromosomal material, rather than gene changes in the strictest sense. The presence of lethal mutants or losses is expressed as a failure of growth, leading to the premature death of the organism carrying them. As most detectable lethal genes are recessive, their effects are observed only in homozygotes. If the processes controlled by a particular gene come early in the developmental sequence, the disturbance usually has fewer ramifications. An analysis of the effects of lethal genes provides a valuable means of investigating the complex processes of embryonic development and cellular differentiation and, in the case of humans, has practical medical implications as well. In microorganisms the study of lethal biochemical mutants opened up a whole new era in the unraveling of biosynthetic processes and led to the establishment of modern biochemical genetics. *See* CHROMOSOME; MOLECULAR BIOLOGY; MUTATION.

Conditional mutants. Although typically lethal genes express themselves independently of internal and external environments, there are cases in which expression is dependent on some particular condition. The temperature-sensitive (*ts*) and amber (*am*) mutants in bacteriophage T4D provide instances where the mutant behaves as a lethal under one set of conditions (restrictive) while under another set of conditions the phenotype of the mutant resembles wild type. Such conditional mutants have been used to analyze gene function in the synthesis and assembly of the macromolecules making up bacteriophage particles. Over 40 conditional lethal mutants have been shown to affect the processes leading to the complete morphogenesis of T4D phage particles. Such conditional lethals are thus invaluable tools for full analysis of gene action in the organization and functioning of biological systems.

Obtaining test materials. One of the difficulties in establishing the mechanism of action of lethal genes in higher organisms is that only one-quarter of the progeny in crosses of heterozygous lethals are homozygous. Analysis at the biochemical level can be greatly simplified if progenies consisting wholly of homozygous lethal individuals are readily obtained. Conditional lethal mutants of the temperature-sensitive type, readily produced in the fly *Drosophila* with the aid of chemical mutagens, which are viable and fertile when raised at 18°C (64°F), and completely lethal at 28°C (82°F), have been the subject of active investigation in several laboratories. By maintaining flies at 18°C and transferring eggs to 28°C, any quantity of lethal tissue becomes available for biochemical study.

Occurrence in natural populations. Lethal genes in natural populations of plants and animals have received considerable attention from geneticists. In certain instances these may form the bases of polymorphisms in such populations. In some organisms they have led to what are known as balanced lethal systems: When two different lethal genes (or losses) are located at different points in different members of a pair of homologous chromosomes in such positions, or circumstances, that recombination by crossing over is prevented, only those individuals heterozygous for both of the two different lethals survive. In this way strains, or even species, which are permanent heterozygotes have arisen as in the evening primroses of the genus *Oenothera*. In human populations the accumulation of lethal mutations have been a matter of some concern in the face of increases in mutagenic agents in the environment and the lowering of selective pressures associated with modern life. The human wastage and misery which accompany high frequencies of lethal genes are considerable. *See* POLYMORPHISM (GENETICS).

Examples. Lethal genes include the following: biochemical lethals in microorganisms such as *Neurospora, Escherichia,* and *Salmonella*; lethal chlorophyll mutants of higher plants; notch and many others in the fruit fly *Drosophila*; creeper in fowl; yellow, brachyury, W-anemia in the mouse; gray-lethal in the rat; Dexter in cattle; thalassemia, sickle-cell anemia, and many others in humans. *See* HUMAN GENETICS.

Donald F. Poulson

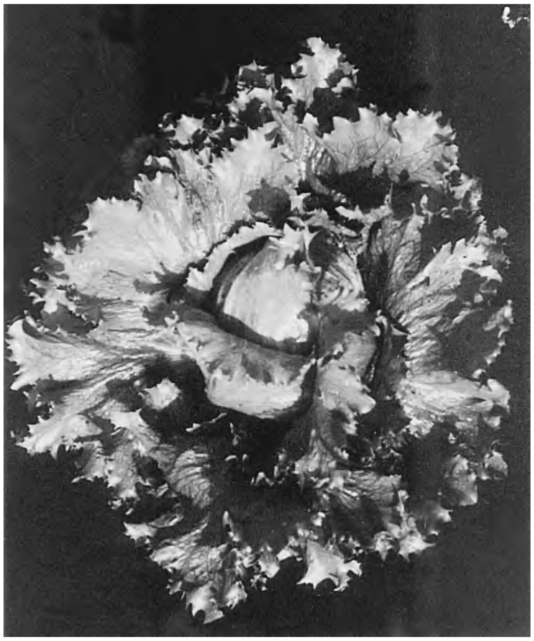
Bibliography. E. Hadorn, *Developmental Genetics and Lethal Factors*, 1961; D. L. Hartl, *Genetics*, 4th ed., 1998; M. M. Moore et al. (eds.), *Mammalian Cell Mutagenesis*, 1988; M. W. Strickberger, *Genetics*, 3d ed., 1985.

Lettuce

A cool-season annual, *Lactuca sativa*, of Asian origin and belonging to the tribe Cichorium of the Compositae family. Lettuce is grown for its succulent leaves, which are eaten raw as a salad. Four varieties of this leading salad crop are head lettuce (*L. sativa* var. *capitata*; see **illus.**), leaf or curled lettuce (*L. sativa* var. *longifolia*), and stem or asparagus lettuce (*L. sativa* var. *asparagina*). There are two types of head lettuce: butterhead, and crisphead or iceberg. *See* ASTERALES.

Propagation. The outdoor crop is propagated by seed usually planted directly in the field, but also occasionally planted in greenhouses for later transplanting. Field spacing varies, but usually plants are grown 10–16 in. (25–40 cm) apart in 14–20-in. (35–50-cm) rows. Greenhouse lettuce, predominantly leaf and butterhead varieties, is transplanted to ground beds with plants placed 7–12 in. (18–30 cm) apart.

Uniformly cool weather promotes maximum yields of high-quality lettuce; 55–65°F (13–18°C) is optimum. Heading varieties (cultivars) are particularly sensitive to adverse environment. High temperatures prevent heading, promote seed-stalk development, and result in bitter flavor and tip-burned leaves. Thus, commercial production of lettuce is extensive in several California and Arizona valleys where mild



Head lettuce (*Lactuca sativa* var. *capitata*). (*Asgrow Seed Co., subsidiary of The Upjohn Co.*)

winter and cool summer climates prevail. Production in other areas with hot summer weather is restricted to the cooler spring and fall seasons.

Crisphead or iceberg lettuce is the most widely grown type; cultivars belonging to the Great Lakes variety account for most of the acreage. Popular cultivars of other types of head lettuce are Butterhead, Dark Green Boston, White Boston, and Bibb; leaf lettuce types are Grand Rapids, Prize Head, and Salad Bowl; cos or romaine types are Parris Island, Dark Green, and Valmaine.

Harvesting. The harvesting of heading varieties begins when the heads have become firm enough to satisfy market demands, usually 60–90 days after planting. Most western-grown lettuce is field-packed in paperboard cartons and chilled by vacuum cooling.

Leaf lettuce and cos lettuce are harvested when full-sized, but before the development of seed stalks or a bitter taste. This varies from 40 to 70 days after planting.

California raises more lettuce than any other state; Arizona, Florida, and Texas are next in importance.

Diseases. Lettuce diseases of greatest importance are caused by viruses, fungi, and bacteria which alter the structure and function of the infected parts, causing stunting, malformation, death, discoloration, and breakdown of tissues.

The most serious fungus diseases (and their causal organisms) are sclerotinia rot or drop (*Sclerotinia sclerotiorum* and *S. minor*), gray mold rot (*Botrytis cinerea*), downy mildew (*Bremia lactucae*), anthracnose (*Marssonina panattoniana*), and bottom rot (*Rhizoctonia solani*).

The most serious of the bacteria parasitic on lettuce are *Erwinia carotovora* and *Pseudomonas marginalis*, both of which cause soft rots. Soft rots are especially damaging when lettuce is not cooled quickly after harvest and is not maintained cool (36–40°F or 2–4°C) during storage and transit to market.

The potentially most serious disease of lettuce is common mosaic, caused by a virus that is seed-borne and transmitted from infected to healthy plants by aphids. Fortunately, this disease can be avoided by use of virus-free seed. Lettuce is susceptible to several other aphid-transmitted viruses that are not seed-borne. Diseases caused by these viruses are usually less prevalent than mosaic and of minor importance.

Important diseases due to other infectious agents are aster yellows, caused by a leafhopper-transmitted spiroplasma, and Big Vein, caused by an infectious entity that has not been seen with the electron microscope and thus has not been identified and characterized. The causal agent is soil-borne and transmitted by a chytrid fungus, *Olpidium brassicae*, that infects the roots of lettuce and many other plants without causing evident injury at the site of infection.

Tip burn, a physiological disease, is especially damaging to head lettuce due to death and browning of leaf tissues inside the head. It is induced by high ambient temperatures occurring near the time of harvest that disrupt normal organic acid and calcium metabolism. Cultivars now being used have been selected for increased tolerance to tip burn. *See* PLANT PATHOLOGY.

R. G. Grogan

Leucite rock

Igneous rocks rich in leucite but lacking or poor in alkali feldspar. Those types with essential alkali feldspar are classed as phonolites, feldspathoidal syenite, and feldspathoidal monzonite. The group includes an extremely wide assortment both chemically and mineralogically.

The rocks are generally dark-colored and aphanitic (not visibly crystalline) types of volcanic origin. They consist principally of pyroxene and leucite and may contain calcic plagioclase or olivine. Types with plagioclase in excess of 10% are called leucite basanite (if olivine is present) and leucite tephrite (if olivine is absent). Types with 10% or less plagioclase are called leucitite (if olivine is absent) and olivine leucitite or leucite basalt (if olivine is present).

The texture is usually porphyritic with large crystals (phenocrysts) of augite and leucite in a very fine-grained or partly glassy matrix. If plagioclase occurs as phenocrysts, it is generally labradorite or bytownite and is slightly more calcic than that of the rock matrix. It may be zoned with more calcic cores surrounded by more sodic margins. Leucite appears in two generations. As large phenocrysts it forms slightly rounded to octagonal grains with abundant tiny inclusions of glass or other minerals zonally arranged. Small, round grains of leucite with tiny glass inclusions also occur in the rock matrix. Augite or diopside (sometimes rimmed with aegirine-augite) and aegirine-augite form the mafic phenocrysts. Pyroxene of the matrix is commonly soda-rich. Olivine may occur as well-formed phenocrysts. Other minerals present may include nepheline, sodalite, biotite, hornblende, and melilite. Accessories include sphene, magnetite, ilmenite, apatite, and perovskite.

Leucite rocks are rare. They occur principally as lava flows and small intrusives (dikes and volcanic plugs). Well known are the leucite rocks of the Roman province, in Italy, and the east-central African province. In the Italian area the feldspathoidal lavas are essentially leucite basalts and may have developed by differentiation of basalt magma (rock melt). In the African province the leucitic rocks are associated with ultramafic (periodotite) rocks and may have been derived from periodotitic material. This may have been accomplished by the abstraction of early formed crystals from a peridotitic magma or by the mobilization of peridotitic rocks by emanations from depth. Assimilation of limestone by basaltic magma may help to decrease the silica content and promote the formation of leucite instead of potash feldspar. The crystallization of leucite, however, is in large part a function of temperature and water content of the magma. Conditions of formation, therefore, may strongly influence the formation of leucitic rocks. *See* IGNEOUS ROCKS; LAVA; PHONOLITE.

Carleton A. Chapman

Leukemia

A disease characterized by a progressive and abnormal accumulation of white blood cells, or leukocytes. Leukemic cells are malignant because they have three characteristics common to all cancers: (1) they exhibit uncontrolled growth that is frequently associated with an inability to mature normally; (2) they arise from a single precursor cell; and (3) they disregard anatomic boundaries and metastasize to organs or tissues where leukocytes are not normally found. The expanding clone of leukemic cells infiltrates organs and tissues, particularly the bloodstream and bone marrow, where they disrupt the production of normal cells. The resulting symptoms include fatigue, pallor, infections, bruising and bleeding, and discomfort caused by enlarged organs. In humans, the term leukemia encompasses more than 20 distinct malignancies. *See* BLOOD; HEMATOPOIESIS.

Diagnosis and classification. Normal leukocytes are grouped into two primary types or lineages, myeloid and lymphoid, and virtually any cell of either lineage can become leukemic. Leukemias are also divided into broad categories that are based on the cell involved (myeloid or lymphoid) and disease aggressiveness (either acute or chronic). Subclassifications are based on morphologic, cytochemical, immunologic, cytogenetic, and molecular criteria. Depending on the type of leukemia and reason for the classification (treatment options, research, or reaching a prognosis), only one or all of the criteria may need to be determined. Some characteristics, including morphologic and cytochemical, can be established within hours or days, whereas others, such as molecular and cytogenetic composition, may take weeks. The length of time involved is important because therapy for acute leukemia usually needs to be started promptly.

Morphologic criteria. A microscopic analysis of stained peripheral blood and bone marrow is the first step in diagnosis (**Fig. 1**). In the appropriate clinical setting, that alone may be all that is needed to classify some leukemias, such as chronic myelocytic leukemia (**Fig. 2**). Morphology can differentiate lymphoid from myeloid types by the presence of Auer rods, which are intracellular structures that appear only in the acute myeloid leukemias (**Fig. 3**).

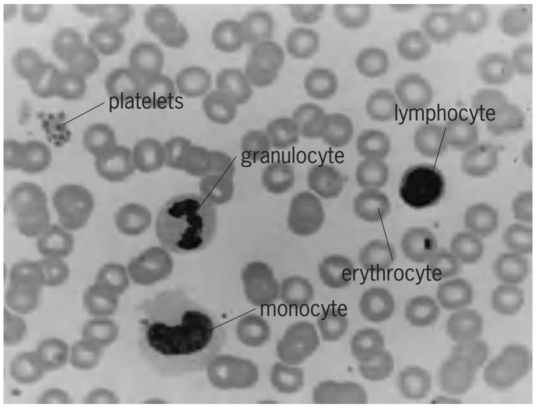


Fig. 1. Normal stained blood smear showing a granulocyte, lymphocyte, monocyte, some erythrocytes, and a clump of platelets.

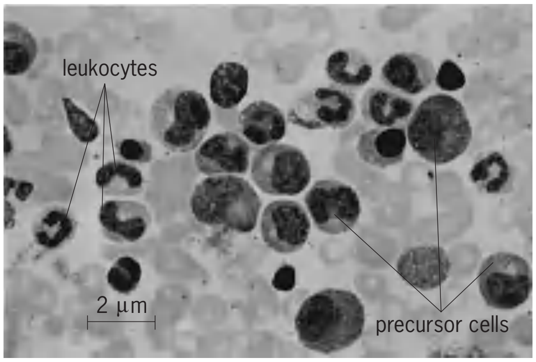


Fig. 2. Stained blood sample taken from a patient with chronic myelocytic leukemia. Many of the cells are bone marrow precursor cells, which are not ordinarily found in the circulating blood. In addition, the total leukocyte count is greatly increased over that found in normal blood.

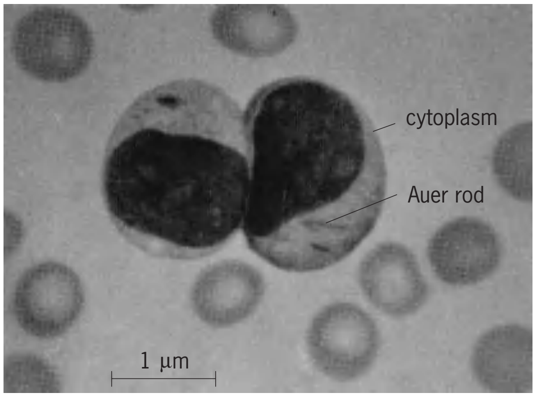


Fig. 3. Two abnormal cells in the peripheral blood of a patient with leukemia, both of which have crystalline structures in their cytoplasm. The crystals, known as Auer rods, are composed of lysosomal enzymes such as acid phosphatase and peroxidase and are virtually always found in patients with acute myeloblastic leukemia.

Cytochemical characteristics. Special stains that impart an easily recognizable color are routinely used to help distinguish cell lineage. One of the most useful stains detects the enzyme peroxidase, which appears only in myeloid cells. Acute myeloblastic leukemias can usually be subclassified by using only morphologic and cytochemical tests.

Immunologic reactivity. Monoclonal antibodies can bind to specific molecules on cells called antigens, and fluorescence or histochemical techniques can be used to reveal antibodies that bind to specific antigens found only on the surface of certain types of cells. Thus, it is possible to assign a leukemia cell to the lymphoid or myeloid categories when the morphologic and cytochemical data are ambiguous or conflicting. Immunodiagnosis has been particularly helpful in subcategorizing lymphoid leukemias and subtyping previously unclassifiable leukemias. Immunologic techniques have also been helpful in establishing clonality (the property of arising from a single cell) and in leading to increased understanding of basic leukemia biology. *See* CELLULAR IMMUNOLOGY; MONOCLONAL ANTIBODIES.

Cytogenetic analysis. Leukemia diagnosis, prognosis, and etiology have been greatly improved through the analysis of individual chromosomes that have been differentially stained to reveal subchromosomal segments. In most leukemias, distinct chromosomal abnormalities appear nonrandomly with much greater frequency in almost all leukemic cells analyzed. The abnormalities are confined to the leukemic cells and are not found in other cells of the body, and so they are acquired defects. The best-known example is the Philadelphia (Ph) chromosome, a translocation of genetic material between chromosomes 9 and 22 that is detected in more than 90% of chronic myeloid leukemia cases. Cytogenetics also provides information on etiology and prognosis. Deletion of part of the long arm of chromosome 5 (5q-syndrome) or of one of the two chromosomes 7 (monosomy 7) strongly suggests that the leukemia was secondary to prior chemotherapy or arose from a preexisting bone marrow disorder, both of which suggest aggressive disease, poor response to therapy, and a shorter survival time. *See* CHROMOSOME; HUMAN GENETICS.

Molecular studies. Molecular biology identifies and analyzes small fragments of deoxyribonucleic acid (DNA) and has helped to confirm the importance of cytogenetics, the analysis of whole chromosomes. Molecular biology has allowed deeper insight into leukemogenesis and the diagnosis and classification of leukemia. For example, if an increased number of lymphocytes are present (lymphocytosis) but the evidence for lineage or malignancy is not clear, a finding of an abnormal amount of specific DNA that encodes for an antigen present only on the surface of T cells (thymus-derived lymphocytes) is a strong indicator of a T-cell leukemia. Molecular biology has been used following apparently curative therapy to screen blood or bone marrow cells for evidence of abnormal cells not detected by other, less sensitive methods such as morphologic or cytochemical analysis. *See* DEOXYRIBONUCLEIC ACID (DNA); MOLECULAR BIOLOGY.

Differentiating leukemia from a normal reaction to some other disease (secondary leukocytosis) is rarely a problem. However, a persistent yet nonmalignant leukocytosis may require all of the previously mentioned techniques to reach a definite diagnosis. *See* CLINICAL PATHOLOGY.

Epidemiology. In the United States an estimated 55,000 cases of leukemia are diagnosed annually and account for 4–5% of all cancers. These figures, and similar ones for other countries, have remained relatively constant since the mid-1960s. In industrialized countries, leukemia is second only to accidents as the most common cause of death among children under the age of 16, and it is the most common malignancy in these countries. Overall, it is the sixth most common cause of death in the United States. Geographic variations in the incidence of leukemias are less apparent than in any other form of cancer. Although worldwide incidence is variable, there is usually no more than a twofold difference between countries or regions within countries. The exception is chronic lymphocytic leukemia, which is the most common form of adult leukemia in western countries but is unusual in the Far East, Africa, and South America. In Japan, chronic lymphocytic leukemia is virtually nonexistent.

Leukemia can occur at any age and even be present at birth, but its frequency rises sharply after the age of 50. Acute leukemias, primarily the lymphoid type, are most common in children and young adults, whereas the chronic leukemias are rare in children. The majority of adult acute leukemias are myeloid. The incidence of chronic myelocytic leukemia peaks between the ages of 40 and 50. Chronic lymphocytic leukemia is rare under age 40; since 90% of cases are diagnosed in people over 50, it is the most common leukemia in the elderly. A slight male predominance is found in all of the leukemias, especially chronic lymphoid, hairy cell, and childhood acute lymphoid leukemias.

Research into environmental or infectious causes has led to the study of clusters or microepidemics of leukemia that are geographically or chronologically contiguous. There are no convincing arguments that leukemia clusters occur other than randomly, except for the adult T-cell form. Although not genetically transmissible in the classic sense, leukemia has a tendency to occur within families. Most striking is susceptibility in identical twins: the risk of the unaffected twin of a sibling with leukemia developing the disease has been estimated at 200 times that of the general population. Risk falls off rapidly for other family members and is only two to three times higher for nontwin siblings. Even a woman who has leukemia and is pregnant rarely if ever gives birth to a child with leukemia. Certain cell surface antigens are found more often in family members of leukemic patients than in the general population, but the search for a "leukemia susceptibility gene" possibly linked to an inherited human leukocyte antigen (HLA) type has been unsuccessful.

Etiology. An estimated 80% of all human cancers are induced by environmental carcinogens, but only a small fraction of exposed individuals develop cancer. Although many agents are suspected of inducing leukemia, for the great majority of cases the etiology is unknown. It appears that no single factor is causative, but a number of events must take place before leukemia occurs.

Radiation. The evidence for ionizing radiation as a leukemogenic cofactor is virtually irrefutable. A dose- and time-dependent relationship with development principally of myeloid leukemias has been shown following radiation exposure from nuclear explosions, with atomic bomb survivors and atomic test participants; from therapy for ankylosing spondylitis (inflammation of the vertebrae) or cervical cancer; and from work-related contact, such as radiology prior to effective shielding. Radiation appears to have no role in the genesis of chronic lymphoid leukemia, and normal exposure to diagnostic x-rays obtained during a lifetime of routine medical care, to natural sources of radiation, or to nonionizing radiation does not appreciably increase the risk of leukemia. *See* RADIATION INJURY (BIOLOGY).

Chemicals and drugs. Chronic exposure to high levels of benzene and perhaps related compounds is associated with a tenfold higher risk of developing myeloid leukemia. The risk from exposure to other chemicals is much less clear, however. Agricultural occupations appear to predispose workers to an increased risk of lymphoma and leukemia, possibly as a result of contact with chemicals or animal-borne viruses. *See* LYMPHOMA.

A clear, strong association has been demonstrated between pharmaceuticals (particularly alkylating agents) that are administered as therapy for a primary cancer and the subsequent development of secondary leukemia, virtually always acute myeloblastic leukemia. Compared with chemotherapy that involves alkylators, radiotherapy alone is regarded as weakly leukemogenic, but the combination of both appears to be worse than either alone. Organ transplantation is usually followed by an increased risk of lymphoma, but leukemia risk also rises. The immune compromise induced by drugs that are used to block graft rejection and possibly the chronic reaction to foreign antigens on the graft as well appear to be related to the induction of malignancies.

Viruses. Only in the rare adult T-cell form of leukemia has a virus been implicated. A strong association has been shown between adult T-cell leukemia and a retrovirus [a type of ribonucleic acid (RNA) virus] called human T-cell leukemia virus I, or HTLV-I. The leukemia that it causes occurs endemically in Japan, especially the islands of Kyushu, Shikoku, and Hokkaido; the Caribbean; the southeastern United States; southern Italy; and parts of South America and Africa. Viruses are known to cause lymphomas and leukemias in nonhuman animals. Examples are bovine and feline leukemia viruses and avian leukosis virus, a retrovirus that is the source of a variety of hematologic neoplasias in domestic fowl. *See* AVIAN LEUKOSIS; FELINE LEUKEMIA; TUMOR VIRUSES; VIRUS.

Congenital chromosomal and acquired abnormalities. Persons with Down syndrome are 30 times more likely to develop acute, usually lymphoid leukemia than the rest of the population. Down syndrome patients have an extra chromosome 21, resulting in a total of 47 chromosomes instead of the normal 46, and the extra chromosome is thought to account for the increased risk for developing leukemia. In fact, leukemias in persons without Down syndrome who have abnormalities of chromosome 21 have been reported. Other rare congenital disorders that are characterized by chromosomal abnormalities and increased risk of leukemia include Fanconi's anemia, a potentially fatal blood disorder; ataxia-telangiectasia, a neuromuscular degenerative disease; and Bloom's syndrome, a condition marked by stunted growth and facial flushing. *See* DOWN SYNDROME.

In the primary immunodeficiency states, malignancies develop 10,000 times faster than in unaffected persons, and each of the immunodeficiencies is associated with a distinct leukemia: X-linked agammaglobulinemia (the condition characterized by lack of or extremely low levels of gamma globulin, together with defective antibody production and frequent infections) with lymphoid leukemia, common variable immunodeficiency with chronic lymphoid leukemia, and ataxia-telangiectasia with acute lymphoid leukemia.

Preexisting hematologic disorders. The myelodysplastic syndromes are characterized by ineffective production of normal blood cells, and result in low blood cell counts due to abnormal precursor cells in the bone marrow. The abnormal cells are clonal and manifest a spectrum of morphologic and cytogenetic abnormalities. Most have a tendency to evolve into acute leukemia, with the myeloid type predominating. Other clonal blood disorders that are characterized by increased production of one specific cell type or by fibrotic marrow have a slight tendency to evolve into acute leukemia. That tendency is greatly enhanced as a result of treatment with alkylating agents or radioactive compounds. A related disorder, chronic myelogenous leukemia, almost always transforms into an acute leukemia, a change referred to as a blast crisis. This process is heralded in 70% of cases by increased myeloblasts and in 25% by proliferation of lymphoblasts or other immature cells. The metamorphosis changes chronic myelogenous leukemia from a relatively well-tolerated chronic condition into a highly lethal leukemia that fails to respond to treatment.

Leukemogenesis. The transformation of a normal cell into a malignant one is complex and not fully understood, but it appears to be caused by a series of events that involve a particular cell in a susceptible host. In all likelihood, these events occur with an exact sequence, and timing and must be of sufficient strength and duration, such as exposure to radiation. The transforming events permanently alter or mutate the genome and confer upon the cell and its progeny a survival advantage over normal cells. Whether there is a block in maturation, as in acute leukemias, or an increase in stem cells (formative cells that have the ability to self-replicate and to give rise to specialized cells) with no maturation defect, such as is seen in chronic myelogenous leukemia, the end result is similar—expansion of the malignant clone.

The importance of cytogenetic changes was made clearer with the use of molecular biologic techniques that led to the discovery of oncogenes. Oncogenes are DNA sequences that were discovered in cancer-promoting viruses and that encode for the production of substances important to the normal cell's transfer of growth signals across the cell membrane. Theoretically, when expressed inappropriately, oncogenes malignantly transform cells. Whether these transformations are by themselves sufficient to cause leukemia is debatable. Cytogenetic and molecular discoveries help to further characterize leukemogenesis but do not yet provide a full explanation. *See* ONCOGENES.

Pathogenesis and clinical course. The two major types of leukemia usually differ in signs and symptoms. Acute leukemias have a relatively rapid onset, and those with the disease often experience problems immediately. Chronic leukemias have an insidious course and are frequently discovered during an examination for an unrelated problem. For both types, the most consistent symptoms are nonspecific and include weakness, fatigue, mild weight loss, and low-grade fever.

Often the leukemic clone first affects blood-forming cells in the bone marrow. To examine the bone marrow, a sample is removed with a needle from the pelvis and then stained and viewed under the microscope. Depending on the type of leukemia, a partial or near-total replacement of normal cells will be apparent. The great numbers of leukemic cells may obliterate all normal bone marrow architecture, including fat spaces that occupy 20% of the marrow cavity in children and as much as 70% in the elderly. It was once thought that decreased production of blood cells was secondary to "crowding out" by the leukemic cells, but it is now clear that the leukemic cells themselves or their products inhibit normal cell production.

Leukemic cells are found in the bloodstream. Microscopic examination of a drop of stained venous blood may reveal only a rare abnormal cell or a vast sea of leukemic cells that number as many as 50 times the normal total white cell count, as in chronic lymphocytic and acute myeloblastic leukemias. Leukemic cells may infiltrate organs and tissues other than the blood and bone marrow and cause disturbances in normal function. Their spread to the brain, meninges, spinal cord, kidneys, intestine, adrenal glands, lungs, and heart may cause pain, loss of function, or obstruction. However, symptoms referable to organs other than the bone marrow usually do not appear until later stages of the disease. The chronic leukemias are exceptions since the symptoms that they often present are referable to an enlarging spleen or liver, and these organs may infarct or rupture as they become progressively engorged with leukemic cells that compromise the blood supply. Another exception occurs in the myeloid leukemias, where tumors are composed of masses of leukemic cells called chloromas. Chloromas contain an enzyme that imparts an evanescent greenish tinge to their cut surface and can precede a diagnosis or herald a relapse of leukemia before it is detectable elsewhere. Chloromas have a predilection for soft tissues and are usually found in or near the sinuses and orbital areas of the skull, vertebrae, lymph nodes, skin, and gonads.

Because the leukemic cell retains most of the physiologic properties of its normal counterpart, the leukemic clone may appear clinically to be a caricature of a function or attribute of the specific cell type from which it originated. Thus, acute progranulocytic leukemia is invariably accompanied by disturbances in clotting and bleeding due to the release of procoagulant substances from granules in the leukemic cells. Hemorrhage can be fatal, particularly after therapy when leukemic cells die and disgorge their contents. Acute monocytic leukemia, which exaggerates the normal monocyte function of egress into sites of inflammation or infection, is often accompanied by a marked degree of organ and tissue infiltration, especially of the gums, spleen, liver, and lymph nodes.

Therapy. Practical therapeutic goals for the acute and chronic leukemias are distinct. Without prompt, intensive, in-hospital therapy, the acute leukemias usually cause death within a few months. In acute leukemia, the object of therapy is to totally obliterate the leukemic clone and allow normal bone marrow cells to recover. This is achieved through large doses of antileukemic drugs that are potent enough to eliminate all evidence of the leukemia. Unfortunately, such agents lack specificity and destroy normal cells particularly the dividing cells in the bone marrow but the doses are calculated to spare normal stem cells that can eventually repopulate the bone marrow and resume normal blood cell production. After the initial intense treatment, repeated cycles are given for as long as 3 years. Specific treatment of localized collections of leukemic cells, especially in so-called sanctuary sites such as the central nervous system, may be included to prevent relapse.

In the chronic leukemias, standard therapeutic principles are completely different. Many patients who initially require no therapy begin mild forms of outpatient treatment as the disease progresses. The intent is not to cure but to control the disease with minimal toxicity. A certain degree of organ compromise (including the bone marrow) is tolerated as long as it is not life-threatening. When therapy is necessary, it usually consists of milder forms of chemotherapy that do not require hospitalization.

Chemotherapeutic agents can be given singly or in regimens that combine two or more drugs. Radiation therapy is used in the preparative phases before bone marrow transplantation, in preventing central nervous system relapse, and in treating focal deposits of leukemic cells.

Chemotherapy. Many effective antileukemic agents have been synthesized. Combination therapy incorporates drugs that have different modes of action and different toxicities in order to increase cytotoxic potency, account for leukemic cells that may be resistant to a single agent, and lessen cumulative toxicity in any particular organ or tissue. Most antileukemic drugs act by perturbing enzymes or substrates that are related to DNA or RNA synthesis and thus largely affect actively dividing cells. Any treatment must be repeated since the number of leukemic cells may exceed one trillion and a single course of antileukemic drugs will destroy only some of them.

Among the first drugs used to treat leukemia were alkylating agents, including cyclophosphamide and chlorambucil, which act by cross-linking DNA and consequently hindering replication and transcription. Similar effects are produced by ionizing radiation. Paradoxically, it is this effect on genetic material that is thought to account for the increased risk of secondary malignancies after exposure to alkylating agents. Anthracyclines, such as doxorubicin, daunorubicin, and idarubicin, inhibit replication by insertion between DNA base pairs and are among the most effective antileukemic agents. Plant alkaloids, such as vincristine and vinblastine, which are derived from the periwinkle, prevent cellular division by inactivating the mitotic apparatus. They are most useful when included as part of a combination regimen, since their mode of action and side effects are different from other commonly used agents. Antimetabolites have been synthesized to be molecular mimics of normal substrates required for processes that are integral to cell survival. They become incorporated into the leukemic cell and, because of chemical substitutions at important sites on the molecule, block essential DNA synthetic pathways and ultimately lead to cell death. The nucleosides pentostatin, fludarabine, lymphoma, and cladribine are most effective in chronic lymphocytic leukemia, lymphoma, and hairy cell leukemia. *See* CHEMOTHERAPY.

Stem cell transplantation. Stem cell transplantation has had the most positive impact on the leukemia cure rate. Chemotherapy is administered alone or with radiation therapy in doses much higher than those used in standard antileukemic regimens to abolish the leukemic clone at the expense of the normal stem cells in the bone marrow. Patients would die following such treatment unless "rescued" with cryopreserved stem cells. The thawed stem cells are infused into a vein to circulate in the bloodstream, occupy the marrow spaces, and eventually proliferate and produce all of the various blood cells. The stem cells must come from the patients themselves (autologous stem cell transplant) or from a donor (allogeneic stem cell transplant) whose human leukocyte antigens, commonly called HLAs, match those of the patient's cells as closely as possible (HLAs determine tissue compatibility). In allogeneic stem cell transplants, stem cells are harvested from the peripheral blood of an identical twin (a syngeneic match) or a sibling or unrelated donor (allogeneic). However, the risk of potentially fatal graft-versus-host disease is significant with donor cells that are less than ideally matched. Many patients do not have eligible siblings and require unrelated matched donors. The elderly, who comprise a large segment of those with leukemias, are generally not candidates because of the toxicity involved. Nonetheless, stem cell transplantation is the therapy of choice for eligible patients with acute lymphoid, acute myeloblastic, and other high-risk leukemias who have relapsed after standard chemotherapy. Allogeneic stem cell transplantation is used for patients with chronic myelogenous leukemia who have failed imatinib treatment (see below) and rarely for patients with chronic lymphocytic leukemia and other leukemias. *See* STEM CELLS.

Biologic response modifiers. The search for therapies which are less toxic and more specific for leukemic cells has focused on substances that are derived from natural (biologic) sources or that affect biologic reactions, some of which are thought to be part of the body's natural defense against cancer. Examples include monoclonal antibodies, small molecules that target protein kinases that affect the cell cycle and lead to tumor cell death, cell products manufactured by recombinant DNA technology, and the patient's own killer cells (large, granular lymphocytes that can lyse a variety of target cells) expanded and activated in the laboratory before reinfusion. Alpha-interferon can be used in certain leukemias, including hairy cell and chronic myelogenous leukemia. Recombinant growth factors, such as granulocyte-macrophage colony stimulating factor to promote leukocyte production, show potential as a therapeutic adjunct, shortening the duration of low cell counts after therapy and possibly augmenting the numbers and activity of cells that can destroy leukemic cells. Rituximab, a chimerized (comprising approximately 33% mouse protein and 67% human protein) anti-CD20 monoclonal antibody, is commonly used in the treatment of B-lymphocyte-derived malignancies, including chronic lymphocytic leukemia. Alemtuzumab is a humanized (comprising 5–10% mouse protein and 90–95% human protein) anti-CD52 monoclonal antibody approved for second-line treatment of chronic lymphocytic leukemia. Gemtuzumab ozogamicin is a humanized anti-CD33 monoclonal antibody conjugated to a potent antitumor antibiotic, calicheamicin, and is approved for the treatment of acute myeloblastic leukemia. Imatinib is a small molecule that targets the BCR-ABL tyrosine kinase unique to chronic myelogenous leukemia and sometimes found in acute leukemia. It has revolutionized the treatment of chronic myelogenous leukemia with over 80% of early diagnosed patients responding to therapy. Dasatinib is an inhibitor of multiple tyrosine kinases and has been approved by the U.S. Food and Drug Administration (FDA) for the treatment of imatinib-resistant patients with chronic myelogenous leukemia. *See* CELLULAR IMMUNOLOGY; MONOCLONAL ANTIBODIES.

Complications and prognosis. Each type of leukemia and every therapeutic modality has unique complications. However, because of bone marrow compromise by therapy or the leukemia itself, the most common problems are low numbers of platelets, red blood cells, and neutrophils. Before the advent of blood-product transfusions, most patients—particularly those with acute leukemia—would die of hemorrhage due to low platelet numbers. Decreased production of red blood cells compounded the anemia from blood loss. Now, however, platelet and red cell transfusions can sustain patients with little to no production of their own blood cells for many months. Transfusion of neutrophils, on the other hand, has not been practical and is rarely used. Without neutrophils, patients become susceptible to infections with organisms that normally live as harmless commensals on the skin and mucous membranes and in the gastrointestinal tract. Antimicrobial therapy has succeeded in treating these infections, but the elimination of resident microbes leaves the leukemic patient open to infections caused by unusual and resistant forms of bacteria, fungi, viruses, and protozoa. Thus, infection has replaced hemorrhage as the leading cause of morbidity and death and overshadows all other complications of therapy. Death from infections usually occurs because therapy fails either to eradicate the leukemia or to promptly restore normal blood cell production. The latter could correct the single most important cause of susceptibility to infections decreased neutrophils. *See* INFECTION.

Despite the differences between the acute and chronic leukemias, and despite the fact that the great majority of patients can be brought into a remission or quiescent phase of the disease, leukemia is one of the most lethal malignancies. If cure is considered to be the absence of disease 3–5 years after cessation of therapy, only a small fraction of all leukemias are curable. An exception is acute lymphoid leukemia in children, where therapeutic advances have resulted in the attainment of a complete remission and cures in the majority of children.

Survival times in other leukemias are different and range from months in some of the acute forms to as long as 20 years in the chronic types. Overall, for adult and childhood acute myeloblastic leukemia, long-term survival is attained in only 20–30% of cases. The figures for adult acute lymphoid leukemia are only slightly better. Stem cell transplantation offers eligible patients a 40–60% chance for a cure. Biologic response modifiers have clearly had an impact on the therapy of many patients with leukemia. *See* CANCER (MEDICINE); ONCOLOGY.

Kenneth A. Foon; Louis Vaickus

Bibliography. D. Catovsky (ed.), *Leukemic Cell*, 2d ed., 1991; J. Fleischer (ed.), *Leukemias*, 1993; A. M. Mauer (ed.), *The Biology of Human Leukemia*, 1990; P. H. Wiernik (ed.), *Neoplastic Diseases of the Blood*, 3d ed., 1995.

Level measurement

The determination of the location of the interface between two fluids, separable by gravity, with respect to a fixed horizontal datum plane. The respective fluids may be any fluids, liquid or gaseous, which do not mix and have specific gravities significantly different from one another. Fluids include granular or particulate solids which are fluidized or handled like fluids. The most common level measurements are, however, between a liquid and a gas or vapour. *See* FLUID; INTERFACE OF PHASES.

Level measurement may be classified in the main categories: direct visual indication of interface location; remotely transmitted indication of interface location; interface location inferred from hydrostatic pressure; interface location inferred from fluid properties. Level measurements may be of an analog or on-off nature.

Direct visual measurement. In many cases, fluid levels may be observed directly and consequently measured to obtain trends or magnitudes in volume.

Graduated scale. Level is measured directly from a vertical graduated scale partially immersed in the liquid.

Glass window. Level is observed through a transparent window in the side of a tank. The window may be graduated with a vertical scale.

Gauge glass. Level is observed in a transparent vertical tube attached to a closed tank. The bottom of the tube is connected to the liquid space and the top of the tube to the gaseous space. The liquid level in the tube corresponds with the level in the tank and may be observed or measured against a graduated scale. Isolating valves usually are fitted in the upper and lower connecting pipes to allow for replacement of the transparent tube (which is usually of glass) without draining the tank (**Fig. 1***a*).

Closed tanks are often under some pressure, hence the use of external tubes able to withstand pressure rather than windows. In pressurized systems the upper and lower connecting tubes are fitted with ball check valves to avoid a dangerous discharge of fluid in the event of a tube rupture. For very high pressure systems, metallic ducts with thick glass windows are used. These windows are sometimes made refractive and artificially illuminated to show more clearly the difference between the two fluids (such as water and steam).

Remote measurement from float. Where levels cannot be observed and hence measured directly, it is common to use a float and to indicate remotely the elevation of this float.

The float must be of an average density between that of the two fluids, the densities of which must be significantly different (such as water and air) to ensure that a sufficient buoyant force is generated on the float, with changing level, to activate the position-sensing mechanism.

The float may generate an analog signal which varies over the whole range of operating level, or may generate an on-off signal as the level rises above or falls below a predetermined elevation. A series of on-off sensors at different elevations can generate a digital type of level indication.

Float with mechanical indicator. Several simple methods allow the float position and hence level to be observed indirectly from outside the vessel containing the fluids. A vertical rod attached to the float and protruding through a hole in the top of the tank can show the level by the amount of protrusion of the rod, which may be graduated. An external weight attached to the float by a rope or tape running over a pulley at the top of the tank can show the level on an inverse scale outside the tank. A rotating shaft passing through a sealed hole in the side of the tank can be connected to the float by a lever so that any rise or fall in the level rotates the shaft appropriately and so moves a pointer on an external graduated scale.

Float with electrical resistance sensor. In a closed vessel a float having a lever connected to a variable-resistance sensor can cause a change in the electrical resistance as the level rises or falls. The change in electrical resistance can be used, via a suitable calibrated electrical instrument, to indicate the level or volume in the tank (Fig. 1*b*).

Float with magnetic switches. In a closed tank a float pivoting about a fixed point can move a magnet close to the wall of the tank down or up as the level rises or falls about a selected level. A similar magnet just outside the tank is flipped by magnetic repulsion to operate electrical contacts to give a corresponding on-off signal. Magnetic repulsion rather than magnetic attraction is employed to create a definite toggle effect (Fig. 1*c*).

Float with buoyancy effect. A float constrained in its vertical movement will exert a varying force on the restraining mechanism. This force in turn can be measured or converted into an analog electrical signal which can be calibrated to indicate the liquid level. In such an application, level can be measured only over the height of the float, so such floats usually have slender dimensions (Fig. 1*d*).

Measurement by hydrostatic pressure. The pressure p at any depth h in a liquid of density ρ is given by the following equation, $p = \rho g h$, where g is acceleration due to gravity. Hence, if the density of liquid is known, the depth of liquid above a selected point can be determined by measuring the pressure in the liquid at that point.

Pressure gauge. In an application with a simple pressure gauge, it may be calibrated to give a direct reading of the depth of liquid. If the tank is closed and there is pressure in the space above the liquid surface, the difference in pressure between this space and the measuring point must be used.

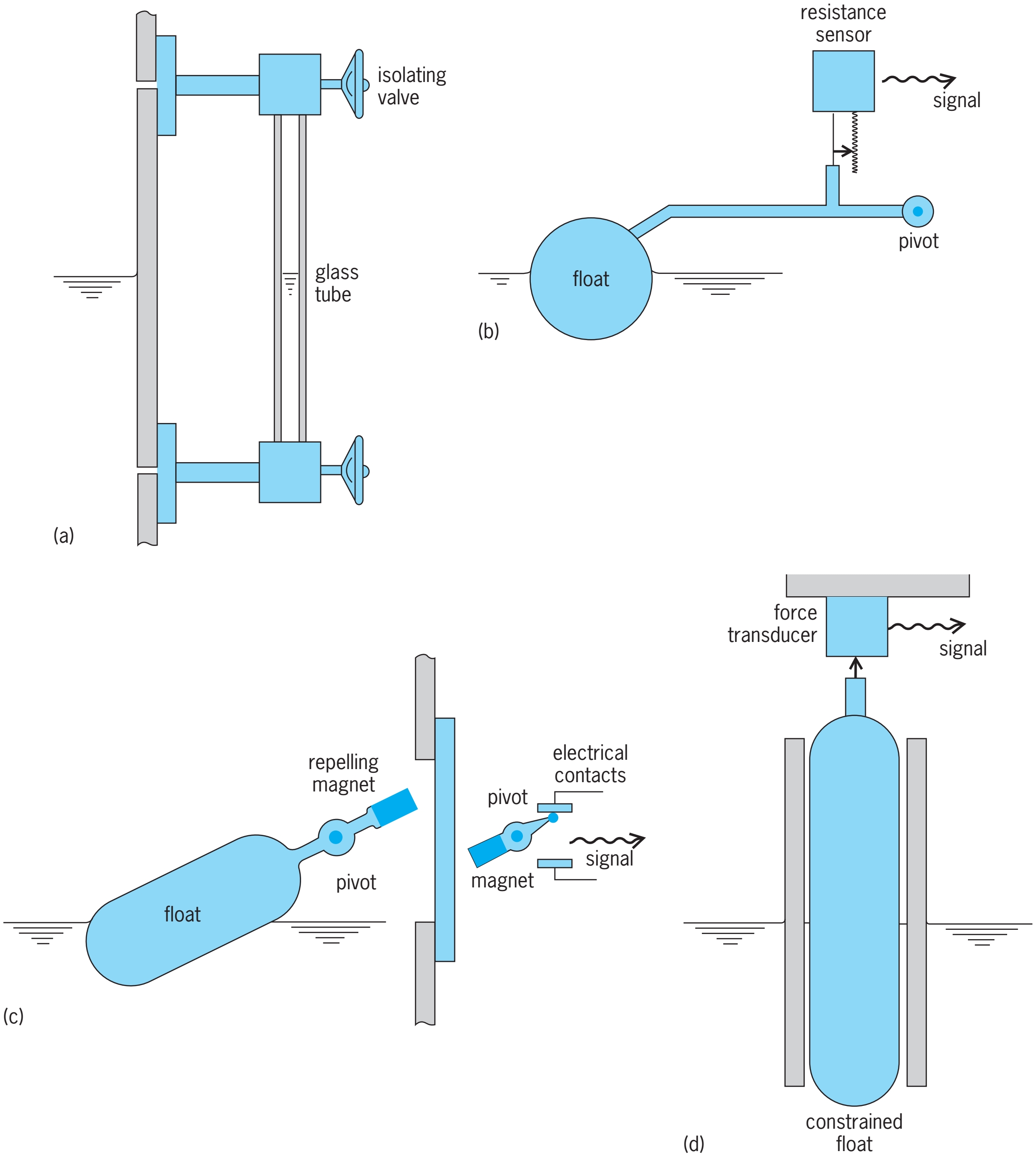


Fig. 1. Direct visual and float-type measurements. (*a*) Gauge glass. (*b*) Float with electrical sensor. (*c*) Float with magnetic switches. (d) Float with buoyancy effect.

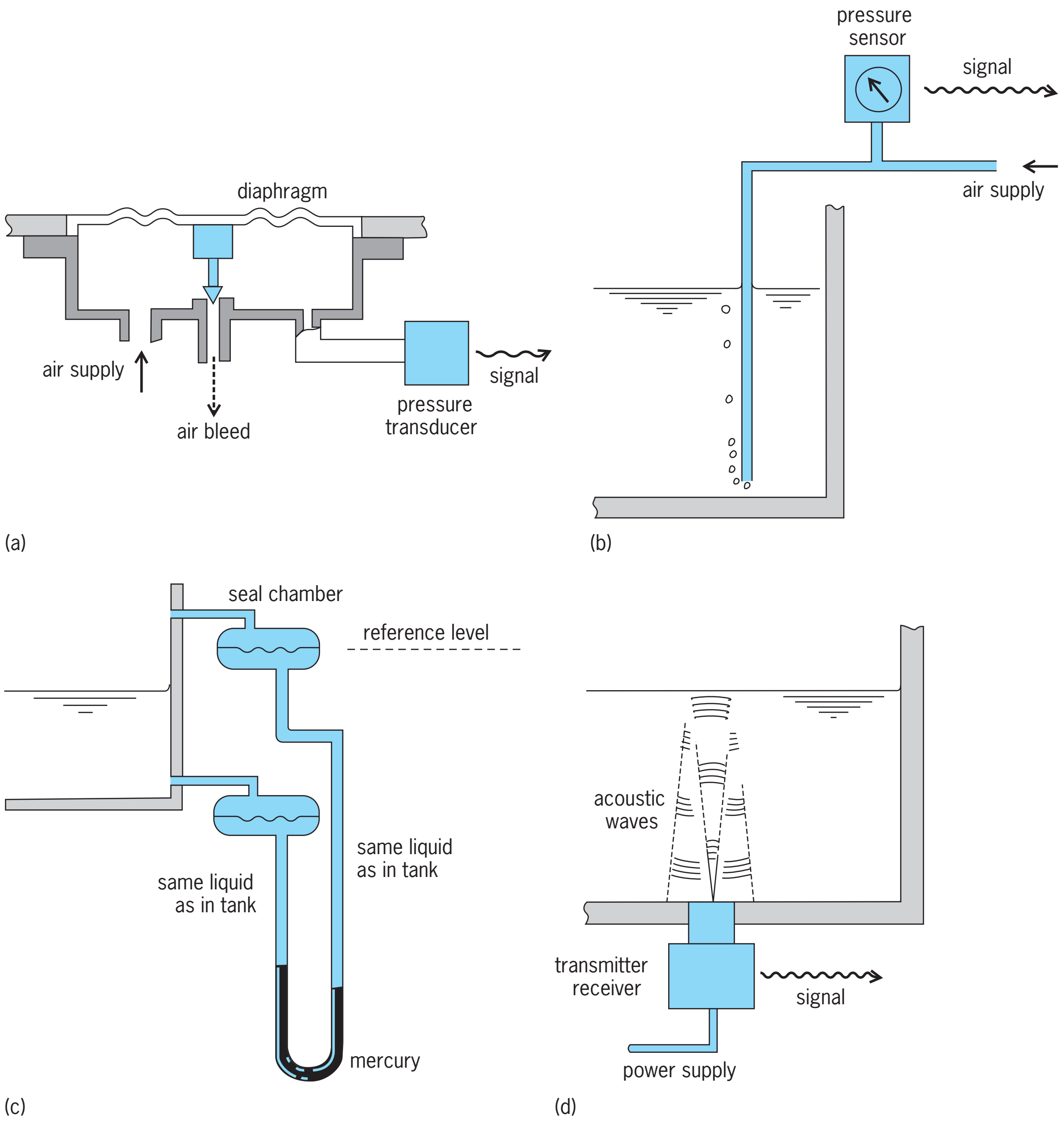


Fig. 2. **Hydrostatic pressure and acoustic-wave measurements. (*a*) Pressure diaphragm. (*b*) Bubble tube. (*c*) Manometer. (*d*) Acoustic.**

Pressure diaphragm. In applications where the liquid may be contaminated with aggressive impurities or contain solids or sludge, pressure gauges or their pressure-measuring tappings may become blocked and unresponsive. In such cases, pressure diaphragms installed flush with the inside surface of the vessel may be used. Movement of the diaphragm against a spring may be measured directly to give an indication of the pressure or depth of liquid. However, since such movement is limited, it is better to apply a corresponding pressure to the back of the diaphragm so as to restore the diaphragm to its neutral position. Measurement of this external pressure then can be converted into a measure of the depth of liquid (**Fig. 2***a*).

Bubble tube. In order to discharge gas from a pipe or vessel into a liquid at some depth, the pressure of the gas must be at least equal to that of the liquid at that depth. Hence the depth of liquid at the point of discharge can be determined from the gas pressure, provided the gas flow rate is low enough to eliminate dynamic or frictional effects. A pipe supplied with a steady but low flow of gas (such as air) may be inserted into a tank of liquid and the gas pressure in the pipe measured. This pressure measurement can then be converted into a measurement of the depth of liquid above the point of discharge (Fig. 2*b*).

Manometer. A manometer may be used instead of a pressure gauge for measuring pressure (more correctly pressure difference). A manometer using mercury as the reference liquid reduces the level variation by a factor of about 13, making direct measurement more convenient, and is more sensitive than a pressure gauge. Its location relative to that of the vessel in which the liquid level is being measured usually necessitates twin pipes from each side of the manometer extending back to the measuring points and filled with the same liquid as in the vessel (Fig. 2*c*).

Measurement by fluid properties. Certain fluid properties can be readily measured or used to determine the presence or extent of a known fluid. The presence or absence of a fluid can provide an on-off signal indicating whether a certain level has been reached or not, whereas the extent of a fluid can be used to determine the depth.

Conductivity. If a liquid is a conductor of electricity, its presence can be detected by a pair of electrodes subject to a potential difference. When immersed on rising level, they can generate an off-on electrical signal.

Capacitance. If a liquid is a dielectric, probes can be inserted into a tank and the capacitance between them measured. This will vary with the degree of immersion and can be converted to a measurement of level. *See* CAPACITANCE.

Acoustic. Most liquids conduct sound waves readily, and these are reflected from any interfaces, including the liquid surface. If an acoustic transmitter-receiver located at the bottom of a tank directs sound waves vertically upward and senses their reflection from the surface, the depth, and hence level, can be determined by the time taken for the sound wave to travel up and be reflected down (Fig. 2*d*). *See* SOUND.

Nuclear. Since gamma rays are absorbed by many liquids, the presence of liquid can be sensed by the attenuation of gamma rays emitted from a gamma-ray source and measured by a detector a short distance away. *See* GAMMA RAYS.

Thermal. If an electrically heated thermistor is subject to immersion in a liquid, it will be cooled more effectively. The resulting drop in temperature will be reflected as a change in resistance which will indicate the presence of the liquid. *See* THERMISTOR.

Robin Chaplin

Bibliography. E. O. Doebelin, *Measurement Systems: Application and Design*, 5th ed., McGraw-Hill, New York, 2003.

Lever

A pivoted rigid bar used to multiply force or motion, sometimes called the lever and fulcrum (**Fig. 1**). The lever uses one of the two conditions for static equilibrium, which is that the summation of moments about any point equals zero. The other condition is that the summation of forces acting in any direction through a point equals zero. *See* INCLINED PLANE.

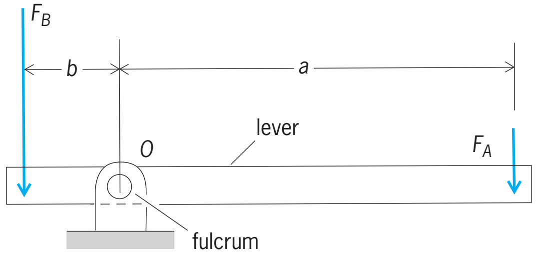


Fig. 1. The lever pivots at the fulcrum.

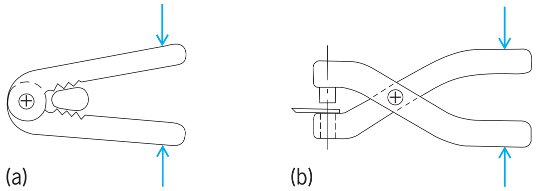

(a) (b)

Fig. 2. Two applications of the lever. (*a*) Nutcracker. (*b*) Paper punch.

If moments acting counterclockwise around the fulcrum of a lever are positive, then, for a frictionless lever, $F_B b - F_{BA} a = 0$, which may be rearranged to give Eq. (1).

$$F_B = \frac{a}{b} F_A \qquad (1)$$

If F_B represents the output and F_A represents the input, the mechanical advantage, MA, is then given by Eq. (2).

$$\mathrm{MA} = \frac{F_B}{F_A} = \frac{a}{b} \qquad (2)$$

Applications of the lever range from the simple nutcracker and paper punch (**Fig. 2**) to complex multiple-lever systems found in scales and in testing machines used in the study of properties of materials. *See* SIMPLE MACHINE.

Richard M. Phelan